Aggregate Outcomes of Nonlinear Prices in Supply Chains[*]

Luca Lorenzini
UCLA Anderson

Antonio Martner
UCLA & Central Bank of Chile

[This version: September 29, 2025]
Link to last version

Abstract

We study the welfare effects of nonlinear pricing in supply chains. Using the universe of firm-to-firm invoices from Chile, we document that unit prices decline with purchase quantity within seller-product-day cells and flatten at higher quantities, with level shifts across buyer industries, patterns consistent with two-part tariffs. We develop a general equilibrium model where firms simultaneously pay and charge nonlinear prices throughout the supply chain. Two-part tariffs push marginal prices toward cost while extracting rents through flat fees which distort firm entry, separating allocation from rent redistribution. Calibrating to the Chilean data, we find that nonlinear pricing achieves 75% of efficient welfare versus 50% under linear pricing. The improvement arises because lower marginal prices reduce markup accumulation across production layers, with entry effects remaining quantitatively modest. These results indicate that, in settings like Chile, nonlinear pricing enhances welfare relative to linear pricing by reducing markup accumulation and that linear pricing assumptions substantially overstate market power costs in supply chains.

[*] We are deeply indebted to Hugo Hopenhayn for his insights and constant encouragement. Antonio would like to thank Michael Rubens and John Asker for invaluable mentorship and support. David Baqaee, Ariel Burstein, Jonathan Vogel, Federico Huneeus, Yasutaka Koike-Mori and Mounu Prem provided highly valuable comments. The views expressed are those of the authors and do not necessarily represent the views of the Central Bank of Chile or its board members. Authors emails: lucalorenzini@ucla.edu, amartner@ucla.edu

1 Introduction

Market power alone does not generate welfare losses: what matters is how that power is exercised through pricing. Under linear pricing, average and marginal prices coincide, so markups summarize quantity distortions. When sellers use nonlinear price schedules, average and marginal prices diverge: marginal prices determine quantity allocations (are allocative) while average prices reflect both allocation and rent redistribution. In such environments, average prices fail as sufficient statistics for welfare. Recent evidence from Bornstein and Peter (2024) shows that nonlinear pricing is pervasive in final-consumption markets and generates welfare losses by misallocating goods across buyers.

In supply chains, nonlinear pricing presents competing forces. Markups compound across production layers, with each layer adding its own markup, generating deadweight losses (the double marginalization problem). Yet nonlinear pricing, particularly two-part tariffs, can reduce the double-marginalization problem by lowering marginal prices toward cost while extracting surplus through flat fees. Whether nonlinear pricing in supply chains improves or reduces aggregate welfare is therefore an empirical question, increasingly relevant given renewed enforcement of quantity discount regulations.[1] This motivates our research question: What are the welfare outcomes of nonlinear pricing in supply chains?

Using universe-level administrative records of firm-to-firm transactions in Chile, we document two facts that contradict the linear pricing assumption. First, for any seller-product pair, unit prices decline with purchase quantity and converge to a common level across buyers, equivalent to two-part tariffs comprising a flat fee and a constant marginal price. Second, under a two-part tariff interpretation, the rate of quantity discounts and the flat fee share vary across seller-industry and buyer-industry pairs, indicating that sellers combine second-degree price discrimination (quantity discounts) with third-degree price discrimination (across buyer groups).

To interpret these patterns and quantify welfare effects, we build a tractable general equilibrium model where firms both charge and pay nonlinear prices throughout the supply chain. The model nests our empirical facts: two-part tariffs separate allocation from rent extraction, with marginal prices determining quantities while flat fees redistribute surplus. This structure distinguishes two welfare channels: an intensive margin, where lower marginal prices reduce markup accumulation relative to linear pricing, and an extensive margin, where rent reallocation affects firm entry. We calibrate all model parameters using the Chilean microdata, then compare equilibria under nonlinear pricing, linear pricing, and efficient pricing. We find that nonlinear pricing achieves approximately three-fourths of efficient welfare, while linear pricing achieves only one-half.

We build on the standard screening framework from mechanism design. We consider a monopolist seller facing heterogeneous buyers with Pareto-distributed types, a distribution that cap-

[1] 2025 FTC actions under the Robinson-Patman Act: FTC v. Southern Glazer's and FTC v. PepsiCo.

tures the empirical regularity of many small buyers and few large ones. When buyer revenue is homothetic in quantities and the seller have constant marginal cost, the optimal nonlinear contract is a two-part tariff comprising a flat fee plus a constant marginal price. The Pareto distribution ensures serving all buyer types is optimal, while homotheticity allows the seller to separate rent extraction from quantity allocation. The marginal price determines quantities, while the flat fee extracts rents without affecting output. This structure delivers a testable prediction: average unit prices decline with quantity and converge to a common marginal price across a seller's buyers. Consequently, average prices mix allocative and redistributive components and fall mechanically with purchase size even when marginal distortions remain constant.

Using Chile's universe of firm-to-firm invoices between formal firms, we find that within each seller-product pair, unit prices fall with purchase size and flatten at higher quantities. This curvature is pervasive across seller industries and shifts with buyer industries, even after absorbing all seller-product-day shocks. The pattern matches the two-part tariff prediction: average unit prices decline over quantity, converging to a common allocative price asymptote across a seller's buyers.

These two facts jointly can rule out pure pricing regimes. A common menu for all buyers (pure second-degree price discrimination) cannot explain price differences across buyers at the same quantity. Linear group pricing (pure third-degree price discrimination) cannot generate the within-seller curvature we observe. The evidence therefore points to a hybrid scheme: sellers use buyer group-specific flat fees and marginal prices, combining third-degree and second-degree discrimination. Since observed quantity-price pairs alone cannot separately identify flat fees from marginal prices, we treat the curvature and level patterns as diagnostic moments that discipline our model and counterfactuals. We reject linear pricing in approximately 70% of transactions.

To quantify the welfare implications of these hybrid pricing patterns, we build a multi-sector general equilibrium supply chain model where firms simultaneously pay and charge nonlinear prices. Homothetic revenue functions allow for arbitrary firm linkages, which, combined with Pareto-distributed buyer types, ensure tractability. Sellers practice hybrid price discrimination: third-degree across buyer sectors and second-degree within sectors. The optimal contract remains a sector-specific two-part tariff with a constant marginal price and a flat fee. This structure nests our empirical patterns and separates allocation from rent extraction at every link in the supply chain. The model yields closed-form expressions for sectoral price indices, costs, allocative markups, profits, and free-entry conditions, with allocative prices varying across buyer-seller sector pairs and remaining below linear allocative prices.

Using sufficient statistics based on final demand exposures, markups, and firm masses, we develop an exact welfare decomposition that separates intensive (allocative) and extensive (variety) margins. This decomposition links sector-level markups to aggregate welfare through exposure weights that trace how final demand depends on upstream costs. We calibrate all model parameters using Chile's universe of firm-to-firm transactions and accounting balance sheets, estimate substitution elasticities using quasi-experimental price shocks, and solve for equilibrium under three scenarios: nonlinear pricing, linear pricing, and efficient marginal cost pricing.

Quantitatively, nonlinear pricing achieves approximately three-fourths of the welfare attainable under efficient pricing, while linear pricing achieves only one-half. Thus, replacing linear with nonlinear pricing recovers half of the lost welfare. We decompose these welfare effects into two channels: an intensive margin (better allocation at existing firms) and an extensive margin (changes in the number of firms). The intensive margin drives most of the improvement, explaining 79% of the welfare gap relative to efficient pricing and 93% under linear pricing respectively. These allocative gains concentrate in upstream sectors that heavily supply final consumers, particularly Retail and Wholesale, Construction, and Manufacturing. Reducing markups in these sectors generates disproportional welfare improvements. The extensive margin effects are smaller: while nonlinear pricing distorts profits and firm firm entry, its aggregate effects are modest compared to the gains from better allocation through lower allocative prices.

This paper delivers a simple message with sharp policy implications. In supply chains, nonlinear pricing can, under certain conditions, improve welfare relative to linear pricing. Our quantitative results using Chilean data show that two-part tariffs push marginal prices toward cost and reduce markup accumulation where final demand is most exposed, recovering a large share of the efficiency lost under linear pricing. The insight for policy debates is that average prices can be misleading: rent extraction through flat fees does not affect marginal prices, and average prices are not allocative. The policy design principle should not be to uniformly ban nonlinear pricing, but rather to assess whether it generates low allocative marginal prices while ensuring rent extraction does not distort firm entry and participation. Our Chilean evidence indicates such welfare-improving conditions can arise in practice.

Our decomposition provides a framework for policy analysis, separating allocative from variety effects and mapping sectoral changes to aggregate welfare through final demand exposure weights. In our Chilean application, this approach identifies where lower marginal prices generate large welfare gains and where rent extraction might distort firm participation. While our specific quantitative findings depend on Chilean market structure and parameters, the methodology itself can be adapted to other settings with sufficient data. This framework could inform empirical evaluation of price discrimination regulations such as Robinson-Patman guidelines, though the welfare implications of nonlinear pricing will likely vary across different market environments and industrial structures.

Related Literature. This paper connects to the literature on firm heterogeneity, market power, and optimal pricing strategies by offering a quantitative framework to explore the macroeconomic effects of nonlinear prices in supply chains. We engage with three strands of literature.

First, we build upon research on price discrimination in intermediate goods markets by documenting its prevalence across firms and quantifying its welfare implications. Our work complements Burstein et al. (2024), who study input price dispersion across buyers and its misallocation effects, by endogenizing firms' pricing behavior and considering both third-degree and second-degree price discrimination. We further examine the welfare consequences of these pricing strategies. Our empirical contribution is to provide new evidence of significant quantity discounts in firm-to-firm transactions. The theoretical contribution is to develop a general equilibrium model in which firms charge and pay nonlinear prices.

Second, we engage with the literature on supply chains by incorporating more pricing mechanisms that better align with empirical patterns. We build on the aggregate market power frameworks of Edmond et al. (2023) and integrate entry dynamics from Baqaee and Farhi (2020), who analyze markup distortions in production networks. By introducing price discrimination, we modify how these distortions propagate through supply chains.

Third, we add to studies on market power by demonstrating how price discrimination can partially mitigate welfare losses from markups through improved resource allocation. While De Loecker et al. (2021) and Boehm et al. (2024) examine the welfare implications of market power in models with firm dynamics, and Hsieh and Klenow (2014) analyze the misallocation effects of resource distortions, we show that incorporating observed pricing practices reduces estimated welfare losses from market power. In this respect, our findings align with Bornstein and Peter (2024), highlighting that price discrimination is a critical factor when assessing the aggregate implications of firm-level market power.

2 Optimal Nonlinear Price Characterization

We present a framework of optimal nonlinear pricing, deriving a lemma that serves as the building block for a general equilibrium supply chain model, in which firms charge and pay nonlinear prices. We find that under Pareto-distributed buyer types, homothetic revenue functions, and constant marginal costs, the optimal price schedule is analytically tractable and can take the form of a two-part tariff: a constant per-unit price and a flat fee.

We use the canonical monopolistic screening problem. A seller with constant marginal cost $c > 0$ faces a continuum of buyers with privately observed productivity types $z \in [\underline{z}, \infty)$, drawn from $F(z)$ with density $f(z)$. The seller chooses a menu, that is, a pair of measurable functions $(q, T) : [\underline{z}, \infty) \to \mathbb{R}_+ \times \mathbb{R}$, which jointly assign to each type z a quantity $q(z)$ and a transfer (i.e. total payment) $T(z)$. Under nonlinear pricing, the transfer $T(z)$ need not equal price times quantity, as under linear pricing, where it would be $p\,q(z)$, and may include, for example, a fixed fee and a per-unit component. Given q, buyer z generates revenue $R(z, q)$ and obtains net surplus $\Pi(z) =$

$R(z,q(z)) - T(z)$. The seller's problem is:

$$\max_{(q,T)} \Pi_{\text{seller}} = \int_{\underline{z}}^{\infty} \Big[T(z) - c\,q(z)\Big] f(z)\,dz \tag{1}$$

$$\begin{aligned} \text{s.t.} \quad (IR) \quad & \Pi(z) \equiv R(z,q(z)) - T(z) \geq 0, \\ (IC) \quad & \Pi(z) \geq R\Big(z,q(\tilde{z})\Big) - T(\tilde{z}), \qquad \forall\, z,\tilde{z} \in [\underline{z},\infty). \end{aligned}$$

Individual rationality (IR) and incentive compatibility (IC) ensure participation and truthful revelation. As shown in Appendix A, the seller's problem can be rewritten as a pointwise optimization:

$$\max_{\{q(z)\}} \Pi_{\text{seller}} = \int_{\underline{z}}^{\infty} \Big[\phi(z,q(z)) - c\,q(z)\Big] f(z)\,dz$$

$$\text{with} \quad \phi(z,q) \;=\; R(z,q) \;-\; \frac{1}{h(z)} \frac{\partial R(z,q)}{\partial z}, \qquad h(z) \;=\; \frac{f(z)}{1-F(z)},$$

where $\phi(z,q)$ is the virtual surplus and $h(z)$ is the hazard rate. The virtual surplus represents the seller's effective revenue from serving type z, and it consists of two components. The first term, $R(z,q)$, is buyer z total revenue. If the monopolist could price discriminate perfectly, this would coincide with the seller's revenue as well. The second term, $-\frac{1}{h(z)} \frac{\partial R(z,q)}{\partial z}$, captures the truth-telling cost, i.e., the additional rents the seller must leave to higher types to prevent them from mimicking type z.

The hazard rate summarizes how many buyers lie above z (i.e., $1 - F(z)$) relative to the density at z (i.e., $f(z)$), and thus measures how easy it is to enforce truth-telling. Taking the FOC for buyer type z:

$$\frac{\partial}{\partial q}\Big[\phi(z,q(z)) - c\,q(z)\Big] = 0 \quad \Longrightarrow \quad \phi_q(z,q(z)) = c, \quad \text{hence:}$$

$$R_q(z,q(z)) \;=\; c \;+\; \frac{1}{h(z)} R_{zq}(z,q(z)) \tag{2}$$

Increasing q for type z yields direct marginal revenue $R_q(z,q(z))$, but it also raises the truth-telling cost by $\frac{1}{h(z)} R_{zq}(z,q(z))$ (the extra rents that must be left to higher types). The optimal contract sets marginal virtual revenue equal to marginal cost.

In addition to constant marginal cost, we impose two further assumptions. First, buyer types are distributed according to a Pareto distribution with tail parameter κ. Second, buyers' revenue functions are homothetic on the quantity transacted with the seller, so buyer type shifts demand for the seller' good without altering its curvature. Specifically, we adopt the normalized homothetic revenue function:

$$R(z,q) = z^{\frac{\sigma-1}{\sigma}}\, q^{\frac{\sigma-1}{\sigma}}, \qquad \sigma > 1,$$

where σ is the curvature parameter (the demand elasticity faced by the seller). These two assumptions impose the tail condition $\kappa > \sigma - 1$ to guarantee finite output demand.

Lemma 1 (Optimal two-part tariff under homothetic revenue and Pareto types[2])**.** *Consider the screening in equation (1). Suppose (i) revenue is homothetic in quantity, with shape parameter $\sigma > 1$; (ii) buyer types are Pareto distributed with tail parameter $\kappa > \sigma - 1$, so that $h(z) = \kappa/z$. Under these assumptions, the optimal nonlinear price schedule is isomorphic to a two-part tariff:*

$$T(z) = F + p^{\mathrm{NLP}}\, q(z), \qquad p^{\mathrm{NLP}} = \frac{\rho}{\rho - 1}\, c, \qquad \rho \equiv \frac{\kappa\, \sigma}{\sigma - 1} > 1, \qquad F:\ \Pi(\underline{z}) = 0.$$

F is a flat fee chosen so that the lowest type's participation constraint binds, $\Pi(\underline{z}) = 0$. We refer to p^{NLP} as the marginal price, to distinguish it from the unit price, since it applies only to the incremental quantity purchased.[3]

Three remarks follow. First, under the Pareto distribution of types, the virtual surplus at the lower bound is strictly positive. Hence, the integrand $[\phi(z, q(z)) - c\, q(z)]\, f(z)$ is positive in a neighborhood of $\underline{z}$, so excluding any mass of low types strictly reduces profit by the foregone positive contribution (see Appendix A.1). Hypothetically, excluding the lowest type $\underline{z}$ would allow the monopolist to raise the flat fee, but at the cost of losing demand from $\underline{z}$. Because a Pareto distribution places a large mass of buyers near the bottom, each excluded buyer contributes little individually, but many are lost at once, making the demand loss larger than any additional flat-fee revenue from those who remain. Therefore, exclusion is never optimal.

Second, the quantities allocated $q(z_i)$ are determined by the marginal price p^{NLP}; the flat fee F only redistributes surplus and does not change q. Seller profit from transacting with type z_i has two components: a variable-profit rectangle $(p^{\mathrm{NLP}} - c)\, q(z_i)$ and a flat-fee component F, which is set by the lowest-type participation constraint (Figure 1a). By the first-order condition in (2), $q(z_i)$ is chosen at the intersection of marginal virtual revenue and marginal cost c; this is implemented by the constant allocative marginal price p^{NLP} and the associated quantity q^{NLP}. The deadweight loss is the area between the demand curve and marginal cost c, over the range of quantities from q^{NLP} to the efficient level q^*. Changing F does not affect this area, while changing the per-unit price does.

Third, the flat fee F is identical across buyer types and purely redistributes surplus. Spreading this fixed amount over more units makes the average unit price fall with quantity: $T(z)/q(z) = p^{\mathrm{NLP}} + F/q(z)$ ((Figure 1b). For small purchases the flat-fee share $F/q(z)$ is large and the average

[2] We get to the same solution using the Wilson (1993) approach on demand profiles as shown in Appendix A.

[3] Proof sketch. Substituting $h(z) = \kappa/z$ and homothetic $R(z, q)$ into the first-order condition (Equation (2)) yields a constant marginal price that solves $R_q(z, q) = \frac{\rho}{\rho-1} c$ with $\rho = \kappa\sigma/(\sigma - 1)$. The envelope condition and IC then pin down transfers up to a constant; choosing F to satisfy $\Pi(\underline{z}) = 0$ completes the two-part tariff. Full derivations are provided in Appendix A.

unit price sits well above the allocative marginal price; as quantity grows, $F/q(z)$ becomes negligible and the average unit price converges to the allocative marginal price p^{NLP}, which governs quantities.

We define the total markup as the ratio of the average unit price to marginal cost. We then decompose this markup into two components. The first component, which we call the allocative markup, is given by the ratio of the marginal price to marginal cost and equals $\frac{\rho}{\rho-1}$. We refer to it as allocative because it alters the quantity allocated, generating in this case a deadweight loss(Figure 1a). The second component is the redistributive markup, which does not affect the allocation but instead redistributes surplus from buyer to seller:

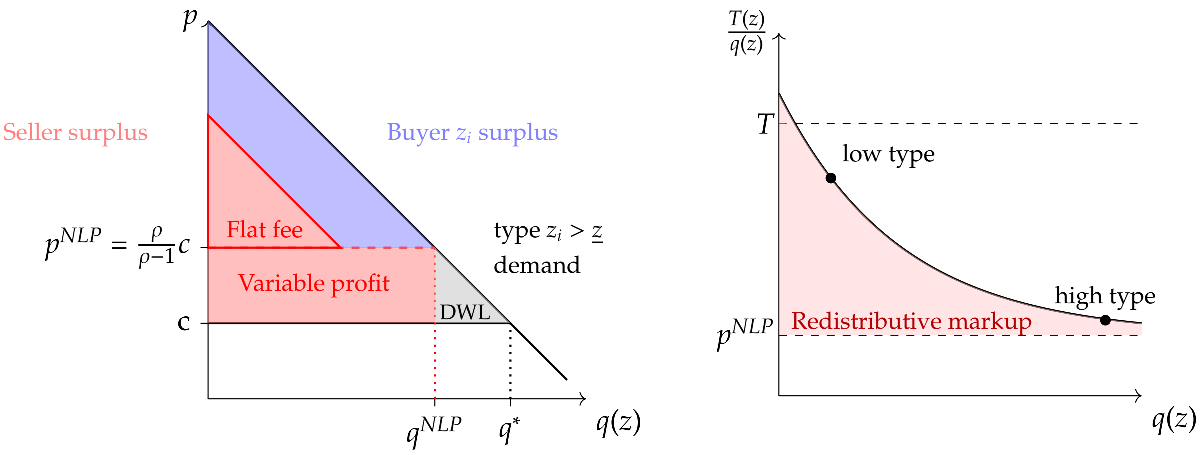
$$\underbrace{\mu_{\text{tot}}}_{\text{Total Markup}} := \frac{T(z)/q(z)}{c} = \underbrace{\frac{\rho}{\rho-1}}_{\text{Allocative Markup}} + \underbrace{\frac{F}{q(z)c}}_{\text{Redistributive Markup}} .$$

It follows that the total markup paid decreases with the quantity purchased, $q(z)$, as illustrated in Figure 1b.

Figure 1: Surplus and Unit Average Prices

(a) Surplus decomposition

(b) Average unit price across types



Note. Panel A: the per-unit price pins down quantity; the flat fee is a lump-sum that redistributes surplus without affecting q. Markup revenue is the rectangle $(p^{\text{NLP}} - c)q^{\text{NLP}}$; efficient and two-part-tariff quantities are labeled q^* and q^{NLP}. Panel B: the average unit price $T(z)/q(z) = F/q(z) + p^{\text{NLP}}$ is higher for low types and declines with z toward p^{NLP}. Distortions in quantities are largest for low types and fade with type.

Implementability in Supply Chains and Testable Footprint. In Section 4, we build a multi-sector environment in which sectors trade with heterogeneous trade intensities. Each seller trades with buyers from multiple sectors, and buyers within each sector are heterogeneous in productivity. Firms both pay nonlinear prices upstream and charge nonlinear prices downstream. As a result, revenue and marginal-cost functions—which in the previous section we treated as primitive—become endogenous, general-equilibrium objects shaped by nonlinear pricing. We assume that sellers can discriminate perfectly across sectors but not across types within a sector. Our main result is that, when firm heterogeneity within a sector follows a Pareto distribution, the equilibrium nonlinear contracts take the form of two-part tariffs, as in Lemma 1. Specifically, we show that the allocative markup and flat fee are seller–sector specific: they are identical across buyers within a sector but vary across sectors.

This characterization delivers a testable footprint: if pricing is isomorphic to a two-part tariff, total payment $T = F + pq$ implies an average unit price $T/q = F/q + p$ that is strictly decreasing and convex in q, with a horizontal asymptote at p (Figure 1b). In the next section, using product-level buyer–seller transaction records from Chile, we test for departures from linear pricing by examining whether average unit prices display this footprint.

3 Evidence on Nonlinear Prices in Supply Chains

We document the presence and shape of nonlinear prices using the universe of firm-to-firm transactions in Chile. Three main findings emerge. First, unit prices vary with quantity transacted and buyer characteristics, inconsistent with linear or uniform pricing. Second, this variation is well-approximated by two-part tariffs: average unit prices fall with quantity while marginal prices converge to a constant. Third, the steepness of these schedules differs across seller and buyer industries, indicating heterogeneity in the strength of nonlinear pricing across the supply chain.

Data Description. We use data from the universe of Chilean firm-to-firm value-added tax invoices collected by the Chilean Internal Revenue Service.[4] For each transaction-specific invoice, we observe seller and buyer IDs, a free-text product "detail," and the corresponding price and quantity. These transaction records can be merged with firms' accounting variables, including total revenue, employee headcounts, labor costs, materials costs, and capital expenditure.

We work at the most granular level and keep the full economy-wide universe of transactions available for 2024, without industry exclusions. Our unit of observation is each invoice line item

[4] This study was developed within the scope of the research agenda conducted by the Central Bank of Chile (CBC) in economic and financial affairs of its competence. The CBC has access to anonymized information from various public and private entities, by virtue of collaboration agreements signed with these institutions. To secure the privacy of workers and firms, the CBC mandates that the development, extraction and publication of the results should not allow the identification, directly or indirectly, of natural or legal persons. Officials of the Central Bank of Chile processed the disaggregated data. All the analysis was implemented by the authors and did not involve nor compromise the Chilean IRS. The information contained in the databases of the Chilean IRS is of a tax nature originating in self-declarations of taxpayers presented to the Service; therefore, the veracity of the data is not the responsibility of the Service

between two tax identifiers.[5] The "detail" field is often seller-specific (e.g., blue paint, brand XX, 3 gallons), so we treat products as seller–product pairs. In most transactions, shipping appears as a separate line, so unit prices exclude shipping.[6] Our approach complements Burstein et al. (2024), who use the same administrative source and document important price-dispersion facts in a manufacturing subsample; here we exploit the complete data available across all industries and retain maximum granularity to study nonlinear pricing in the aggregate supply chain.

We perform three minimal data-cleaning steps to limit measurement error. First, we keep transactions with positive prices and quantities and non-missing product detail. Second, we keep firms that reported positive sales in at least one month during 2024. Third, to avoid spurious variance, we drop products with at least two transactions where the same-day max–to–min price ratio exceeds the 99th percentile of its daily distribution. These filters retain 98% of transactions. The final sample contains 537,521 seller IDs and 3,398,323 buyer IDs that traded 60,029,741 distinct products across 1.24 billion transactions in 2024.

Price Determinants. We begin by quantifying within–seller–product price dispersion. Following Burstein et al. (2024), we construct normalized prices at two frequencies. For each seller–product–month cell (i, g, m) with at least two transactions, define $\tilde{p}_{ijgm} \equiv \frac{p_{ijgt}}{\bar{p}_{igm}}$ where p_{ijgt} is the unit price charged by seller i to buyer j for product g at time t and $\bar{p}_{igm}$ is the mean unit price for (i, g) in month m. Analogously, for each seller–product–day cell (i, g, d) we define $\tilde{p}_{ijgd} \equiv p_{ijgt}/\bar{p}_{igd}$. Residual price variation at the monthly level may reflect inflation or supply shocks, using daily residuals mitigates this concern.

The variance of $\ln \tilde{p}$ is 0.65 at the monthly frequency and 0.61 at the daily frequency; about 29% of cells display no within–cell dispersion (all transactions occur at a single unit price). Histograms are reported in Appendix B.1. These facts reject uniform pricing within seller–product pairs and motivate a decomposition of the residual dispersion into quantity versus buyer components. To net out common shocks, we first estimate:

$$\ln p_{igjt} = \beta_0 + \Psi_{igd} + \epsilon_{ijgt},$$

where Ψ_{igd} are seller×product×day fixed effects and t indexes the time stamp during the day. The residual ϵ_{ijgt} captures price differences across buyers of the same (i, g) on the same day. We then explain ϵ_{ijgt} using alternative fixed–effect sets S:

$$\epsilon_{ijgt} = \beta_0 + \Psi_S + \nu_{ijgt}, \tag{3}$$

[5] Not necessarily firms as some tax IDs do not report hiring workers, purchasing intermediate inputs, or capital expenditure.

[6] Including shipping could generate declining average unit prices with quantity, which would reflect scale economies in shipping rather than nonlinear pricing contracts. By excluding shipping charges, we ensure that variation in average unit prices reflects contractual form rather than transportation technology.

where Ψ_S includes (i) functions of transaction quantity, (ii) buyer–group fixed effects (sector × size × region; 526 groups), and (iii) interactions of quantity with buyer–group to allow group–specific discount schedules. This two-step procedure yields an R^2–based horse race over residual price variation. The results provide indicative evidence on the importance of second-degree (quantity) and third-degree (buyer) components. Identification nuances in separating these mechanisms, and additional robustness using monthly fixed effects and industry subsamples, are detailed in Appendix B.2.

Table 1 reports the R^2 across specifications. Quantity alone explains 34% of residual price variation (Column 1), indicating that quantity discounts play an important role in explaining residual price variation. Coarser buyer–group effects (sector × size × region) still account for 28% (Column 2), indicating that most of the variance explained by buyer fixed effects is captured by observable group characteristics. Allowing group–specific discount schedules (Quantity × Buyer–group) explains 53% (Column 3), consistent with hybrid second– and third–degree price discrimination accounting for the lion's share of price dispersion along supply chains.[7]

Table 1: Price residual determinants

	(1)	(2)	(4)
R^2	0.344	0.275	0.535
S = Quantity	✓		
S = Buyer Group		✓	
S = Quantity × Buyer group			✓
N	147M	147M	147M

Notes: The table reports R^2 values from regressions of price residuals ϵ_{ijgt} on different specifications S, where residuals are obtained from equation 3 after controlling for seller-product-day fixed effects. Buyer groups are defined by combinations of 11 sectors, 3 size categories, and 16 regions.

Nonlinear Prices. We test for nonlinear pricing by examining whether observed (equilibrium) unit prices covary systematically with transaction quantities. We estimate

$$\ln p_{igjt} \;=\; \beta_1 \ln q_{igjt} \;+\; \Psi_{igd} \;+\; \Psi_S \;+\; \varepsilon_{igjt}, \tag{4}$$

where p_{igjt} and q_{igjt} are the unit price and quantity for seller i, product g, buyer j, at transaction day t on day d. Ψ_{igd} are seller×product×day fixed effects; Ψ_S varies by specification to add buyer or buyer–group controls and their interactions (Table 2). A potential concern for interpreting β_1 as evidence of price discrimination is that supply shocks could simultaneously reduce prices

[7] We omitted buyer fixed effects alone because of absence of price variation by buyer and day for the same seller and product. Appendix B.2 includes them an the monthly level indicating stable heterogeneity across buyers, but never generating a higher R^2 relative to quantities and buyer groups interacted.

and raise quantities, creating spurious correlation. Because we condition on seller×product×day fixed effects, identification comes only from within-seller-product-day price variation, making this interpretation unlikely. Another concern is that some buyers may systematically purchase larger quantities and also obtain lower prices due to monopsony power; by including buyer or buyer–group controls in Ψ_S, we assess how much of the observed variation can be explained by this mechanism. We estimate (4) on the universe of 2024 transactions after dropping singletons.

Table 2: Price-quantity coefficient estimates

	(1)	(2)	(3)	(4)
$\ln q_{igjt}$	-0.042	-0.084	-0.065	-0.037
	(0.0001)	(0.0001)	(0.0001)	(0.0001)
S_{Base} = Seller × Product × Day	✓			
$S = S_{Base}$+ Buyer		✓		
$S = S_{Base}$+ Buyer Group			✓	
$S = S_{Base}$ × Buyer Group				✓
N	430M	430M	430M	430M
R^2	0.9646	0.9678	0.9659	0.9790

Notes: The table reports coefficients from regressions of log unit prices on log quantities with varying fixed effect specifications S. Base refers to seller × product × day fixed effects. Buyer groups are defined by combinations of 11 sectors, 3 size categories, and 16 regions. Standard errors in parentheses. All regressions use the universe of Chilean firm-to-firm transactions in 2024 after dropping singletons.

Column (1) conditions on seller×product×day and yields a quantity coefficient of −0.042, so doubling quantity is associated with a 4.2% lower unit price. Adding buyer fixed effects in Column (2) strengthens the coefficient to −0.084, indicating that once persistent buyer heterogeneity is absorbed, quantity discounts are even more pronounced. Replacing buyer FE with buyer–group FE (sector×size×region) still gives a sizable −0.065 in Column (3). Column (4) allows fully flexible group-specific schedules by interacting Ψ_{igd} with buyer group; the coefficient remains negative and precisely estimated at −0.037—about 90% of the Column (1) magnitude—consistent with systematic quantity discounts across buyer groups.[8]

We repeat the same exercise from column (1) for each 1-digit sector in the economy and show the results in Appendix B.3. We find that the smallest quantity coefficient is around 0% in utilities while the largest is observed in construction at 13%.

Could buyer bargaining power drive the price–quantity correlation? Table 2 argues against it: adding buyer fixed effects strengthens the coeficient from −0.042 to −0.084, whereas a buyer–power story would predict attenuation once persistent buyer heterogeneity is absorbed. As a second

[8] We do not interact quantity with buyer fixed effects. Within a day for a given seller–product, the same buyer rarely purchases multiple distinct quantities; moreover, such a specification would push toward buyer-specific nonlinearities closer to first-degree discrimination, which we view as implausible in this setting.

check, we proxy buyer power by the number of distinct suppliers a buyer transacts with and interact $\ln q_{igjt}$ with this proxy. Reassuringly, the interaction is precisely estimated at an economically negligible magnitude; full results are in Appendix B.4. Taken together, the evidence points to seller-side nonlinear pricing rather than buyer bargaining power as the primary driver of the observed patterns.

Nonlinear pricing via within-product quantity quantiles. Because products trade at very different scales, we compare prices across ranks in each product's quantity distribution rather than raw quantities. For each product g, let $F_g(\cdot)$ be the empirical CDF of transacted quantities q_{igjt} using all 2024 observations of product g, and define the within-product rank:

$$r_{igjt} \;\equiv\; F_g\big(q_{igjt}\big).$$

Partition $[0,1]$ into 50 equal-probability intervals $I_b \equiv \big((b-1)/50,\ b/50\big]$ for $b = 1, \ldots, 50$, and assign each transaction to a bin $B_{igjt} = b$ whenever $r_{igjt} \in I_b$.[9] We then estimate the saturated specification:

$$\ln p_{igjt} \;=\; \beta_0 \;+\; \sum_{b=2}^{50} \beta_b \, \mathbb{1}\big\{B_{igjt} = b\big\} \;+\; \Psi_{igd} \;+\; \varepsilon_{igjt}, \tag{5}$$

where Ψ_{igd} are seller×product×day fixed effects and bin b=1 (smallest purchases) is the omitted category. By construction, bin b represents the same position in each product's quantity distribution, making the schedule comparable across heterogeneous products traded in heterogeneous units while absorbing all (i, g, d) shocks. Thus, identification of the coefficient comes solely from within–seller–product–day price variation.

Figure 2 summarizes the estimated schedule. Panel A plots the coefficients $\{\beta_b\}$ from (5) (with bin b=1 omitted) and, for readability, reports percent discounts relative to the smallest-quantity bin as $\Delta_b \equiv 1 - \exp(\beta_b)$. Prices fall steeply over the lower ranks: by b=10, unit prices are about 15% lower than in b=1. Discounts then continue to deepen but at a slower rate, stabilizing around 18–20% for mid-to-large purchases. The final bin shows an additional dip, consistent with products or relationships concentrated in bulk trades (e.g., pack sizes or contract lots). The shape—convex at small quantities with a clear flattening at higher ranks—is consistent with nonlinear pricing under a two-part tariff, where unit prices converge to a constant.

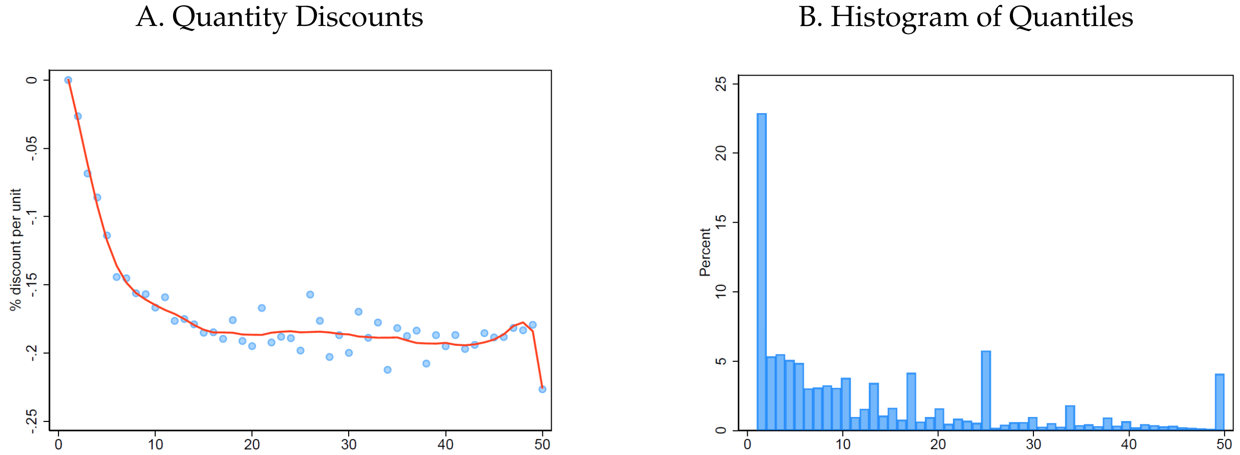
Price discrimination through a two-part tariff with flat fee F and constant marginal price p, as in Lemma 1, yields the average unit price $\bar{p}(q) = p + F/q$. Hence unit prices fall steeply at small q and flatten as q grows, approaching the constant marginal price p from above. The empirical schedule—steep discounts at low quantities and clear flattening at higher quantities—matches

[9] With discrete quantities and mass points, we assign observations to the smallest b such that $r_{igjt} \in I_b$;. Products with fewer than 50 distinct ranks are handled by the empirical CDF

the prediction of Lemma 1, providing evidence consistent with a two-part tariff as the optimal nonlinear pricing strategy. Heterogeneity in F or p across buyer groups (i.e., from third-degree price discrimination) shifts the curves vertically without altering their shape, while discrete pack sizes or volume thresholds can generate the extra dip in the top bin by spreading F over unusually large q.

Panel B validates the rank-binning: despite highly skewed raw quantities, the distribution of observations across bins is close to uniform, with expected excess mass at the extremes due to common integer/mass-point quantities (e.g., single units) and bulk packs. This confirms that the schedule compares positions in product-specific quantity distributions rather than raw scales, allowing aggregation across heterogeneous products while absorbing (i, g, d) shocks.

Figure 2: Prices by Quantity Quantiles



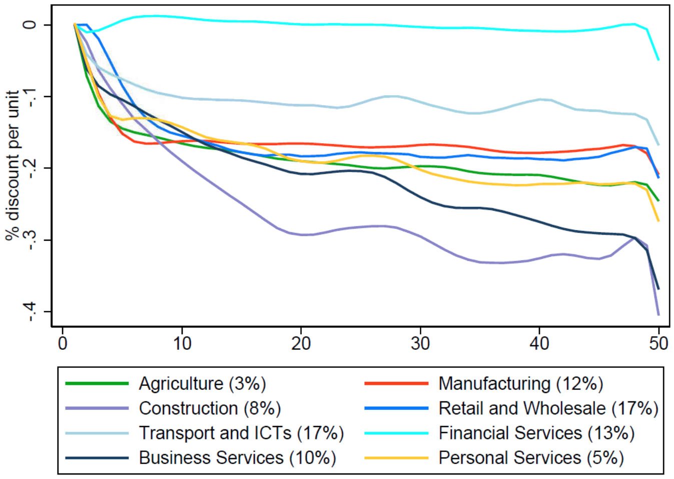
A. Quantity Discounts

B. Histogram of Quantiles

Notes: This figure summarizes the nonlinear relationship between quantity and price. Panel A plots the estimated coefficients from regression (5), where log unit prices are regressed on 50 product-level quantity quantile indicators, controlling for seller-product-day fixed effects. The red line represents a fitted local polynomial of degree 5. Panel B shows the distribution of observations across the modified quantiles, illustrating the re-binning strategy where each unique quantity is consistently mapped to a quantile across products. The first and last bins are overrepresented due to mass points in single and bulk purchases.

Heterogeneity Across Seller Industries. We re-estimate equation (5) separately by 1-digit seller industry to compare price schedules across sectors. Figure 3 shows pronounced between-sector heterogeneity in both steepness and curvature. Business Services and Construction exhibit the largest declines—cumulative discounts approaching 35–40% by the top ranks—while Manufacturing and Retail & Wholesale display moderate but clear discounts (roughly 15–20%). Transport and ICT are comparatively flat, and Financial Services is nearly flat across the entire rank distribution. Despite level differences, the qualitative shape (steep at low ranks, flattening at high ranks) is common, consistent with two-part tariffs where the fixed component is more salient for small purchases. In the model we develop in Section 4, this between-sector heterogeneity arises endogenously from differences in industry competition and in sellers' ability to appropriate buyer surplus.

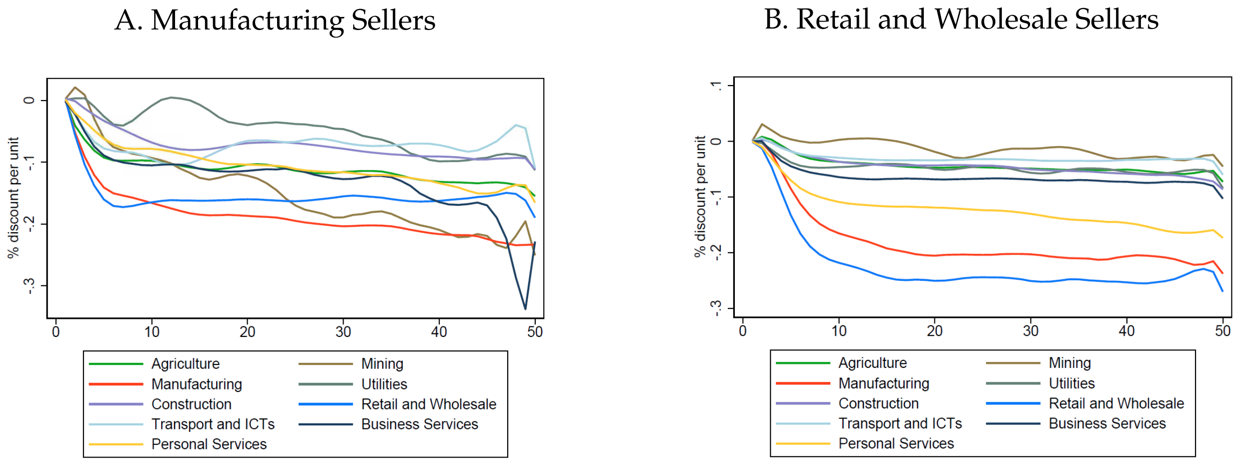
Figure 3: Prices by Quantity Quantiles, by Seller Industry



Notes: This figure plots quantity discount schedules estimated separately by 1-digit seller industry. Each line represents a fifth-degree polynomial fit to the 50 fixed effects estimated from Equation (5), where the dependent variable is log unit price and the main regressor is a quantile bin of quantity, with seller-product-day fixed effects included. The y-axis measures the percent discount per unit relative to the lowest-quantity transactions. The x-axis denotes the quantity quantile bin, ranging from 1 (smallest purchases) to 50 (largest). Sector labels include each industry's share of total GDP (excluding exports) in parentheses.

Buyer-Industry Heterogeneity Within Seller Sectors. We next fix a seller industry and re-estimate Equation (5) separately by 1-digit buyer industry, recovering buyer-sector–specific schedules within each seller sector. Figure 4 illustrates the two largest seller sectors by number of product transacted in 2024. For Manufacturing sellers (Panel A), discounts are markedly steeper for buyers in Manufacturing, Mining, and Business Services—unit prices fall by roughly 15–20% from the lowest to highest ranks—while schedules for Utilities and Transport & ICT are comparatively flat. For Retail & Wholesale sellers (Panel B), buyers in Manufacturing, Retail & Wholesale, and Personal Services exhibit sizable declines, whereas other buyer sectors display near-flat profiles (at most ≈5% even at the top ranks). Because all specifications include seller×product×day fixed effects, these patterns reflect differential within-day, within-product price–rank relationships by buyer type. The evidence is consistent with a hybrid of second- and third-degree price discrimination: sellers deploy nonlinear schedules but tailor their menus to observable buyer characteristics.

Figure 4: Prices by Quantity Quantiles, by Buyer Industry (Within Seller Sector)



Notes: We fix the seller industry and estimate Equation (5) separately for each 1-digit buyer industry. The fitted lines represent fifth-degree polynomials of the estimated quantity-bin fixed effects. Each curve corresponds to a specific buyer sector and traces the percent discount per unit relative to the smallest purchases. The x-axis represents quantity quantiles from 1 (smallest) to 50 (largest).

Taking Stock. Within seller×product×day cells, unit prices decline with quantity ranks and flatten at higher ranks. This curvature is pervasive across seller industries and shifts systematically with buyer industries (Figures 3–4). Because all (i, g, d) shocks are absorbed, these patterns reflect within-seller-product-day price differences that reject uniform pricing and are consistent with a hybrid of second- and third-degree price discrimination: second-degree screening drives curvature, while observable buyer type shifts levels and steepness across industries. Guided by these facts, the next section develops a multi-sector supply-chain general equilibrium model with heterogeneous firms where contracts feature second and third-degree price discrimination.

4 A Model of Nonlinear Prices in Supply Chains

We develop and characterize a general-equilibrium supply-chain model in which firms simultaneously act as buyers and sellers, and therefore both pay and charge nonlinear prices. We show that the optimal contract takes the form of a two-part tariff—a constant marginal price combined with a fixed fee, extending Lemma 1 to the general-equilibrium setting. The framework provides closed-form sufficient statistics for welfare analysis under counterfactual policies. We consider two such counterfactuals: (i) an economy restricted to linear pricing, corresponding to a ban on price discrimination; and (ii) the planner's allocation, which coincides with the decentralized equilibrium when price discrimination is banned and output is subsidized.

4.1 Environment and Notation

There are two firm types, $\ell \in \{u, r\}$, defined by their position with respect to final demand. Upstream firms (type u) sell to retailers and other upstream firms, and source inputs from upstream suppliers. Retailers (type r) purchase inputs from upstream firms and sell exclusively to the representative final consumer.[10] There is a finite set of sectors $\mathcal{S}$, common to both firm types. We use $s \in \mathcal{S}$ for buyer sectors and $s' \in \mathcal{S}$ for seller sectors. In each (ℓ, s), there is a continuum of firms that differ in their productivity z. Productivity z is Pareto distributed within (ℓ, s) with lower bound $\underline{z}_s^\ell > 0$ and tail parameter $\kappa_s^\ell > 0$; the support is $z \in [\underline{z}_s^\ell, \infty)$.

When a firm appears as a buyer we index it by i, and when it appears as a seller we index it by j.The mass of firms in (ℓ, s) is N_s^ℓ. We treat N_s^ℓ as endogenous under free entry in later sections. We work in a steady state and omit time subscripts for brevity.

Market Structure. Retailers sell to the representative consumer at uniform per-unit prices,[11] while sourcing inputs from upstream firms at nonlinear prices. Upstream firms likewise purchase inputs from other upstream firms at nonlinear prices and sell their own variety to both retailers and other upstream firms. Consistent with the evidence we find for Chile, sellers j observe the buyer's type and sector pair (ℓ, s) but not the idiosyncratic buyer productivity z_i; they know only its (Pareto) distribution. They set type and sector-specific tariff schedules but cannot condition on z_i, implying third-degree price discrimination across (ℓ, s) and second-degree within each (ℓ, s).[12]

Preferences. The representative consumer owns all firms and inelastically supplies one unit of labor (L=1). Let P_Y be the final–goods price index.[13] Final demand is Cobb–Douglas across retail sectors with within–sector CES over retail varieties:

$$Y = \prod_{s \in \mathcal{S}} Y_s^{\theta_s}, \qquad \sum_{s \in \mathcal{S}} \theta_s = 1, \tag{6}$$

$$Y_s = \left(\int_{j \in \mathcal{R}_s} y_j^{\frac{\varphi_s - 1}{\varphi_s}} \, d\nu_s(j) \right)^{\frac{\varphi_s}{\varphi_s - 1}}, \tag{7}$$

where $\theta_s \in (0, 1)$ are Cobb–Douglas output elasticities, $\varphi_s > 1$ is the within-sector elasticity of substitution, and $\mathcal{R}_s$ is the set of active retail sellers in sector s. Here $d\nu_s(j)$ denotes the equilibrium measure over active retail sellers $j \in \mathcal{R}_s$, with total mass $N_s^r \equiv \nu_s(\mathcal{R}_s)$.

[10] This two-type structure is motivated by Chilean administrative data showing that most firms sell either only to final consumers or only to other firms, with minimal overlap; see Appendix B.5.

[11] For welfare effects of nonlinear pricing on final demand, see Bornstein and Peter (2024).

[12] Sustaining type and sector-specific tariff schedules requires the absence of zero-cost arbitrage or resale; secondary markets may fail to emerge due to repackaging costs, regulation, or other frictions.

[13] If we normalize $P_Y \equiv 1$, welfare equals real final expenditure Y.

Technology. Firms (buyers i) produce with Cobb–Douglas technology in labor and a Cobb–Douglas aggregator across seller sectors; denote the buyer sector by s and the seller sector by s':

$$Q_i = z_i \, l_i^{\alpha_s^\ell} \, M_i^{1-\alpha_s^\ell}, \qquad 0 < \alpha_s^\ell < 1, \tag{8}$$

$$M_i = \prod_{s' \in \mathcal{S}} M_{is'}^{\theta_{ss'}^\ell}, \qquad \sum_{s' \in \mathcal{S}} \theta_{ss'}^\ell = 1 \quad \text{for each } (\ell, s), \tag{9}$$

where Q_i is output, z_i is firm i's productivity, l_i is firm-level labor input, α_s^ℓ is the labor output elasticity for firms of type ℓ in sector s, and M_i is the composite materials bundle. $M_{is'}$ is the materials bundle from upstream seller sector s', and $\theta_{ss'}^\ell \geq 0$ are input elasticities for buyers in (ℓ, s) across seller sectors $s' \in \mathcal{S}$.[14]

Within any seller sector s', the materials bundle is CES across firm varieties with elasticity $\sigma_{s'} > 1$ for each $s' \in \mathcal{S}$:

$$M_{is'} \;=\; \left(\int_{j \in \mathcal{U}_{s'}} m_{ij}^{\frac{\sigma_{s'}-1}{\sigma_{s'}}} \, d\nu_{s'}(j) \right)^{\frac{\sigma_{s'}}{\sigma_{s'}-1}}, \tag{10}$$

where m_{ij} is buyer i's input of seller variety j in seller sector s', $\sigma_{s'}$ is the elasticity of substitution across those varieties, and $d\nu_{s'}(j)$ is the equilibrium measure over active upstream sellers $\mathcal{U}_{s'}$, with total mass $N_{s'}^u \equiv \nu_{s'}(\mathcal{U}_{s'})$.

Input Price–Taking. Firms are atomistic in input markets and take the wage as given. They retain market power in output markets due to product differentiation under CES demand.

Firm Entry. Firm entry follows Hopenhayn (1992) and Melitz (2003), adapted to a supply-chain environment. In each (ℓ, s) there is an unbounded pool of identical potential entrants. Entry requires paying a sunk cost $c_s^{E\,\ell} > 0$ in units of labor, after which firms draw their productivity z. Active firms exit exogenously at the end of the period with probability $\delta_s^\ell \in (0, 1]$, which serves as the only source of time discounting.[15] Let $\pi^{\ell s}(z)$ denote a potential entrant's per-period profit in numeraire units. Free entry requires that the expected discounted value of profits equals the entry cost in every (ℓ, s):

$$\frac{1}{1-\delta_s^\ell} \, \mathbb{E}_z \left[\pi^{\ell s}(z) \right] \;=\; c_s^{E\,\ell} \, w, \quad \forall (\ell, s),$$

[14] A zero weight $\theta_{ss'}^\ell = 0$ means sector s as a buyer does not use inputs from sector s'. Under no price discrimination and linear prices, $\{\theta_{ss'}^\ell\}$ coincide with input–output expenditure shares for buyer sector s, as in Acemoglu et al. (2012).

[15] Because we focus on steady-state comparisons of macroeconomic outcomes and abstract from time discounting aside from δ, the model is isomorphic to either a constant z over time or a stochastic process for z under the counterfactual of interest.

where the expectation is taken over the post-entry distribution of z in (ℓ, s).

Market Clearing. All markets clear in equilibrium. Labor market clearing requires that the total demand for labor across all active firms equals the inelastic supply of one unit:

$$\sum_{\ell \in \{u,r\}} \sum_{s \in \mathcal{S}} \int_{i \in \mathcal{F}_{\ell s}} l_i \, d\nu_{\ell s}(i) \;=\; L \;=\; 1,$$

where $\mathcal{F}_{\ell s}$ is the set of active firms of type ℓ in sector s, and $\nu_{\ell s}$ is the equilibrium measure over these firms.

For each upstream variety $j \in \mathcal{U}_{s'}$, market clearing requires that output equals the sum of inputs demanded by all buyers. For each retail variety $j \in \mathcal{R}_j$, market clearing requires that output equals final demand from the representative consumer:

$$Q_j \;=\; \sum_{\ell \in \{u,r\}} \sum_{s \in \mathcal{S}} \int_{i \in \mathcal{F}_{\ell s}} m_{ij} \, d\nu_{\ell s}(i), \quad \forall j \in \mathcal{U}_{s'},\ s' \in \mathcal{S}; \qquad Q_j \;=\; y_j \quad \forall s \in \mathcal{S},\ \forall j \in \mathcal{R}_s.$$

General Equilibrium under Nonlinear Pricing. Within a period: (i) potential entrants in each (ℓ, s) pay $c_s^{E\,\ell}$ and then draw productivity z; (ii) each upstream seller $j \in \mathcal{U}_{s'}$ observes only the buyer's pair (ℓ, s) (not z_i) and offers a pair–specific nonlinear contract menu $\{m_j^{\ell,s},\ T_j^{\ell,s}\}$; retail sellers $j \in \mathcal{R}_s$ post linear prices to final consumers; (iii) buyers $i = (\ell, s, z_i)$ observe the offered menus and the wage w and choose labor l_i and input bundles $\{m_{ij}\}_j$ to maximize profits; (iv) production and trade occur, transfers $\{T_{ij}\}_j$ are realized, and final demand $\{y_j\}$ is met; (v) firms exit with probability δ_s^{ℓ}. Contracts are enforceable, resale/arbitrage is ruled out, and beliefs are rational; we consider a steady state so all aggregates are time-invariant.

A general equilibrium consists of allocations $\{Q_i,\ l_i,\ \{m_{ij}\}_j\}$, transfers $\{T_{ij}\}_j$, and consumer demands $\{y_j\}$ for all buyers $i = (\ell, s, z_i)$ with $\ell \in \{u, r\}$ and $s \in \mathcal{S}$, such that: (i) technologies (8)–(10) hold for every firm; (ii) each upstream seller $j \in \mathcal{U}_{s'}$ chooses contracts $\{m_{ij}, T_{ij}\}_i$ that solve its profit–maximization problem (defined below), while each retail seller $j \in \mathcal{R}_s$ sets linear prices to final consumers; (iii) each buyer chooses labor l_i and input bundles $\{m_{ij}\}_j$ to maximize profits given the wage w and the offered contracts or prices; (iv) retail market clearing: $Q_j = y_j$ for all $s \in \mathcal{S}$ and all $j \in \mathcal{R}_s$; (v) upstream market clearing: $Q_j = \sum_{\ell \in \{u,r\}} \sum_{s \in \mathcal{S}} \int_{i \in \mathcal{F}_{\ell s}} m_{ij} \, d\nu_{\ell s}(i)$ for all $s' \in \mathcal{S}$ and all $j \in \mathcal{U}_{s'}$; (vi) the labor market clears; and (vii) free entry holds in each (ℓ, s).[16] A proof of existence and uniqueness is provided in Appendix C.7.

[16] Policy counterfactual equilibrium is defined analogously, except that firm maximization problems are subject to the additional constraints implied by the policy experiment (e.g. a ban on price discrimination or the introduction of output subsidies).

4.2 Guesses: Contracts and Revenue Shapes

We characterize the equilibrium using a guess-and-verify approach. In a supply-chain setting, both firm costs and revenues are shaped by price discrimination. We begin by positing functional forms for contracts that deliver a tractable marginal cost function, which allows us to analyze the firm's unrestricted price-discrimination problem. We then verify the conjecture by showing that the resulting equilibrium coefficients are internally consistent.

Motivated by Lemma 1, we conjecture that optimal contracts are isomorphic to a two-part tariff specific to (ℓ, s). Furthermore, we conjecture that revenue functions are homogeneous of degree ψ_s^ℓ in output.

Guess 1: Two–Part Tariffs by Buyer Type and Sector (ℓ, s)**.** For a buyer $i = (\ell, s, z_i)$, the total payment for purchasing seller variety $j \in \mathcal{U}_{s'}$ is conjectured to take the form of a two–part tariff, with the marginal price determined by an (ℓ, s)–specific markup $\mu_{ss'}^\ell$:

$$T_{ij} \;=\; p_{js}^\ell \, m_{ij} \;+\; F_{js'}^\ell \;=\; \mu_{ss'}^\ell \, c_j \, m_{ij} \;+\; F_{js'}^\ell,$$

where m_{ij} is the quantity purchased by buyer i from seller j, $p_{js}^\ell = \mu_s^\ell \, c_j$ is the marginal (allocative) price, ℓ and s denote the buyer's type and sector, and c_j is the seller's marginal cost. The fixed component $F_{js'}^\ell$ varies with the seller identity j and with the buyer only through the observable pair (ℓ, s').

Guess 2: Equilibrium Buyer Revenue Function. We conjecture that, in equilibrium, the revenue function is homogeneous of degree ψ_s^ℓ in output:

$$R_i \;=\; A_s^\ell \left(Q_i\right)^{\psi_s^\ell},$$

for parameters A_s^ℓ and ψ_s^ℓ that are constant at the buyer's type–sector (ℓ, s) level.

4.3 Preliminaries

To proceed, it is useful to derive input demand, price indices, and cost functions under the conjectured contract structure. Since flat fees are inframarginal, they do not affect these objects but only redistribute profits across firms. As a result, input demand, price indices, and cost functions coincide with those in a linear-pricing economy, except that prices vary at the (ℓ, s) level. We then solve the unrestricted price-discrimination problem for a generic seller and verify the conjecture by matching undetermined coefficients.

4.3.1 Costs and Price Indices

Using the guesses in Section 4.2, in particular, that marginal prices are quantity–invariant within a buyer type–sector (ℓ, s) and a seller sector s', we can define sectoral price indices and derive firm costs. These objects will be verified once we solve for equilibrium prices.

CES Sectoral Price Index. For any seller sector $s' \in \mathcal{S}$ and buyer type–sector (ℓ, s), let p^{ℓ}_{js} denote the marginal price charged by seller variety $j \in \mathcal{U}_{s'}$ to buyers in (ℓ, s). With elasticity $\sigma_{s'} > 1$, the unit price of the s'-bundle faced by buyers in (ℓ, s) is:

$$P^{\ell}_{ss'} = \left(\int_{j \in \mathcal{U}_{s'}} \left(p^{\ell}_{js} \right)^{1-\sigma_{s'}} d\nu_{s'}(j) \right)^{\frac{1}{1-\sigma_{s'}}}, \tag{11}$$

where $d\nu_{s'}(j)$ is the equilibrium measure over sellers in s', and $N^{u}_{s'} \equiv \nu_{s'}(\mathcal{U}_{s'})$ denotes their total mass. Flat fees do not enter (11).

Cobb–Douglas Materials Cost Index. For firm $i = (\ell, s, z_i)$, the unit price of its composite materials bundle M_i in (9) is:

$$P^{M}_{i} = \prod_{s' \in \mathcal{S}} \left(P^{\ell}_{ss'} \right)^{\theta^{\ell}_{ss'}}, \qquad \sum_{s' \in \mathcal{S}} \theta^{\ell}_{ss'} = 1, \;\; \theta^{\ell}_{ss'} \geq 0. \tag{12}$$

Firm–Level Marginal Cost. Only marginal prices $\{p^{\ell}_{js}\}$ enter via (11)–(12); transfers T_{ij} are inframarginal and do not affect marginal cost. Given technology, wage $w > 0$, and constant returns to scale, the marginal cost of producing Q_i units for firm $i = (\ell, s, z_i)$ is:

$$c_i = \frac{\Theta^{\ell}_{s}}{z_i} w^{\alpha^{\ell}_{s}} \left(P^{M}_{i} \right)^{1-\alpha^{\ell}_{s}}, \qquad \text{where} \quad \Theta^{\ell}_{s} \equiv \left(\alpha^{\ell}_{s} \right)^{-\alpha^{\ell}_{s}} \left(1 - \alpha^{\ell}_{s} \right)^{-(1-\alpha^{\ell}_{s})} \prod_{s' \in \mathcal{S}} \left(\theta^{\ell}_{ss'} \right)^{-(1-\alpha^{\ell}_{s})\theta^{\ell}_{ss'}}.$$

Sectoral Productivity Index. Following Melitz (2003), define the CES sectoral productivity index for upstream (seller) sector s' and retail (seller) sector s as:

$$\widetilde{z}^{u}_{s'} = \left(\int_{j \in \mathcal{U}_{s'}} z_j^{\sigma_{s'}-1} \frac{d\nu_{s'}(j)}{N^{u}_{s'}} \right)^{\frac{1}{\sigma_{s'}-1}}, \qquad \widetilde{z}^{r}_{s} = \left(\int_{j \in \mathcal{R}_{s}} z_j^{\varphi_{s}-1} \frac{d\nu_{s}(j)}{N^{r}_{s}} \right)^{\frac{1}{\varphi_{s}-1}}.$$

4.3.2 Buyer Input Demand

Given the guesses in Section 4.2 and the objects defined in Section 4.3.1, each buyer $i = (\ell, s, z_i)$ chooses labor and input quantities from upstream seller varieties to maximize profits. Flat fees are infra–marginal and do not affect marginal conditions; only marginal prices p_{is} matter for input choices. For notational simplicity, we therefore formulate the maximization problem in terms of profits $\widetilde{\Pi}_i$, net of flat fees.

Let m_{ij} denote the quantity of seller variety $j \in \mathcal{U}_{s'}$ purchased from seller sector s'. Using Guess 2, buyer $i = (\ell, s, z_i)$ solves:

$$\widetilde{\Pi}_i = \max_{l_i,\, \{m_{ij}\}_j} \left\{ A_s^\ell \, Q_i^{\psi_s^\ell} \; - \; w \, l_i \; - \; \sum_{s' \in \mathcal{S}} \int_{j \in \mathcal{U}_{s'}} p_{js}^\ell \, m_{ij} \, d\nu_{s'}(j) \right\}.$$

The total expenditure on inputs from seller sector s' can be expressed as $P_{ss'}^\ell M_{is'}$. The first–order condition with respect to $M_{is'}$ equates the marginal revenue product of the materials bundle to its sectoral price index, similarly for labor:

$$\frac{\partial R_i(z_i, \{M_{is'}\}, l_i)}{\partial M_{is'}} \; = \; P_{ss'}^\ell, \qquad \frac{\partial R_i(z_i, \{M_{is'}\}, l_i)}{\partial l_i} \; = \; w.$$

which determines the labor–materials ratio given $\{P_{ss'}^\ell\}$. This input demand implies that the marginal revenue product of the materials bundle from sector s' is equalized across firm varieties within (ℓ, s). For a given buyer i, demand for the materials bundle from upstream sector s', denoted $M_{is'}$, is allocated across varieties $j \in \mathcal{U}_{s'}$ according to the CES share rule. We have that p_{js}^ℓ is the price charged to buyers in (ℓ, s) and $P_{ss'}^\ell$ is the sectoral price index in (11). Under the conjecture that markups are (ℓ, s)–specific, it implies that relative input demands across varieties depend only on a seller's marginal cost relative to the sectoral price index. Buyer identity enters solely through the scale term $M_{is'}$.

$$m_{ij} \; = \; M_{is'} \left(\frac{p_{js}^\ell}{P_{ss'}^\ell} \right)^{-\sigma_{s'}} \; = \; M_{is'} \left(\frac{\widetilde{z}_{s'}^u}{z_j \, (N_{s'}^u)^{\frac{1}{1-\sigma_{s'}}}} \right)^{-\sigma_{s'}}, \qquad \sigma_{s'} > 1,$$

where $\widetilde{z}_{s'}^u$ denotes the productivity index for sector s' and $N_{s'}^u$ is the measure of active upstream sellers in s'. Using this condition, together with market clearing, we can express total production of variety j as a function of its relative productivity:

$$Q_j \; = \; \left(\frac{z_j}{\widetilde{z}_{s'}^u} \right)^{\sigma_{s'}} \; Q_{s'}\big(\widetilde{z}_{s'}^u\big), \tag{13}$$

where $Q_{s'}(\widetilde{z}_{s'}^u)$ denotes the total production of the average firm in upstream sector s'.

Scaling of Input Demand with Productivity. Input usage scales with firm productivity relative to the sectoral average.[17] For an upstream firm $j \in \mathcal{U}_{s'}$ with productivity z_j, labor and material demand satisfy:

$$l_j(z_j) = \left(\frac{z_j}{\widetilde{z}^u_{s'}}\right)^{\sigma_{s'}-1} l_j\left(\widetilde{z}^u_{s'}\right), \qquad M_j(z_j) = \left(\frac{z_j}{\widetilde{z}^u_{s'}}\right)^{\sigma_{s'}-1} M_j\left(\widetilde{z}^u_{s'}\right).$$

Hence, if productivities are Pareto–distributed with tail parameter $\kappa_{s'}$, input demand is also Pareto with different tail parameter for upstream firms and retailers based on their relevant elasticity of substitution:[18]

$$\xi^u_{s'} = \frac{\kappa^u_{s'}}{\sigma_{s'} - 1}, \qquad \xi^r_s = \frac{\kappa^r_s}{\varphi_s - 1} > 1.$$

4.4 The Optimal Nonlinear Price

Fix an upstream seller sector s' and a buyer type–sector (ℓ, s). A seller $j \in \mathcal{U}_{s'}$ offers an unrestricted menu $\{x, T\}$ to buyers $i = (\ell, s, z_i)$, where x denotes the allocated quantity and T the transfer.

To describe the buyer i surplus and the extractable rents by seller j when transacting with i, let $\nu_{s'}$ denote the equilibrium measure over upstream sellers in sector s'. Denote buyer i's profit (inclusive of transfers) by Π_i. For buyer i, the total surplus from transacting with seller j of productivity z_j is defined as:

$$TS_{is'}(m_{ij}) := \left.\frac{d\,\Pi_i}{d\big(\nu_{s'}(z_j)\big)}\right|_{\arg\max \Pi_i},$$

namely, the marginal value to buyer i of access to an additional infinitesimal mass of sellers of type z_j within sector s', evaluated at buyer i optimal input choices. This is the surplus that an infinitesimal seller j seeks to appropriate through its contract.

We can express this is surplus in terms the marginal revenue product. Under CES aggregation within sector u' (elasticity $\sigma_{u'} > 1$), the extractable surplus from a purchase of a generic size m can be written as:

$$TS_{is'}(m) = \frac{\sigma_{s'}}{\sigma_{s'} - 1} \frac{\partial R_i(z_i, \{M_{is'}\}, l_i)}{\partial M_{is'}} M_{is'}^{\frac{1}{\sigma_{s'}}} m^{\frac{\sigma_{s'}-1}{\sigma_{s'}}} - T,$$

where m is the quantity purchased and T the associated transfer. Since the individual seller is infinitesimal, it treats the marginal revenue product as parametric and given. Using the first–order condition, this simplifies to:

$$TS_{iq}(m) = \frac{\sigma_q}{\sigma_q - 1} P^{\ell}_{jq} M_{iq}^{\frac{1}{\sigma_q}} m^{\frac{\sigma_q-1}{\sigma_q}} - T.$$

[17] This follows from the homogeneity of the Cobb–Douglas technology: $y(x_1, \ldots) = x_1 \cdot y(1, x_2/x_1, \ldots)$ implies that input ratios are pinned down by common input prices. Scaling by relative productivity yields $\frac{l(z)}{l(\widetilde{z})} = (\frac{z}{\widetilde{z}})^{\sigma-1}$, so input demand inherits a Pareto distribution with effective tail parameter $\xi = \kappa/(\sigma - 1)$.

[18] Equilibrium feasibility requires $\xi^u_{s'} > 1$ and $\xi^r_s > 1$. Aggregate labor demand is $L = \int l(z)\, d\nu(z)$, which is finite only if $\xi > 1$; hence $\kappa^u_{s'} > \sigma_{s'} - 1$ upstream and $\kappa^r_s > \varphi_s - 1$ in retail.

Valuation index and its distribution. For a seller in sector s', the buyer i matters only through the one–dimensional valuation index:

$$\tau_{is'} \;\equiv\; P^{\ell}_{ss'} M^{1/\sigma_{s'}}_{is'}.$$

The valuation index, $\tau_{is'}$ combines the sector–s' price level $P^{\ell}_{ss'}$ with the buyer's scale $M_{is'}$ in the exact way that determines the marginal revenue from a small purchase: from the CES share rule, a seller's marginal revenue is proportional to $\tau_{is'}\, m^{(\sigma_{s'}-1)/\sigma_{s'}}$ for quantity m (up to the transfer T). Hence, a seller's problem can be written solely in terms of $\tau_{is'}$.

When buyer productivities z_i in (ℓ, s) are Pareto, $\tau_{is'}$ is Pareto as well. Let $M_{is'}$ be the buyer's materials demand from sector s'; under our technology, $M_{is'}$ is strictly increasing in z_i. Therefore $\tau_{is'}$ inherits the Pareto law of the productivity distribution z, with:

$$\tau_{is'} \;\sim\; \text{Pareto}\Big(\rho^{\ell}_{ss'}\Big), \qquad \rho^{\ell}_{ss'} \;=\; \sigma_{s'}\, \xi^{\ell}_{s}.$$

Rescaling by $P^{\ell}_{ss'}$ shifts only the scale (not the tail) of the distribution. Feasibility requires $\xi^{\ell}_{s} > 1$, which implies $\rho^{\ell}_{ss'} > \sigma_{s'}$ for all (ℓ, s, s').

Seller's Problem. Applying the revelation principle, a seller in sector s' chooses menus of allocations $x(\tau)$ and transfers $T(\tau)$ for all buyers $i = (\ell, s, z_i)$. Total profit maximization problem is:

$$\max_{\{x(\cdot),\,T(\cdot)\}} \;\sum_{\ell\in\{u,r\}} \;\sum_{s\in\mathcal{S}} N^{\ell}_{s}\, \mathbb{E}_{\tau_{is'}} \Big[\, T(\tau) - c_j\, x(\tau) \,\Big], \tag{14}$$

subject to:

$$\begin{aligned} \text{(LIC)} \quad & TS^{\ell}_{ss'}(\tau) = TS^{\ell}_{ss'}(\underline{\tau}) + \frac{\sigma_{s'}}{\sigma_{s'}-1} \int_{\underline{\tau}}^{\tau} x(\omega)^{\frac{\sigma_{s'}-1}{\sigma_{s'}}}\, d\omega, \\ \text{(IR)} \quad & TS^{\ell}_{ss'}(\underline{\tau}) \;\geq\; 0, \\ \text{(Monotonicity)} \quad & x(\tau') \;\geq\; x(\tau) \quad \text{for } \tau' > \tau. \end{aligned}$$

The local incentive-compatibility constraint (LIC) ensures that truth-telling is optimal for all buyer types τ locally; (IR) is the individual-rationality constraint, binding for the lowest type $\underline{\tau}$; and monotonicity requires that higher buyer types receive weakly larger allocations. Taken together, (LIC) and monotonicity guarantee that the mechanism is incentive-compatible globally. Because the objective and constraints are additively separable across (ℓ, s, s') triples, the problem can be solved independently for each partition.

Solution concept. The seller's mechanism design problem is isomorphic to the setting of Lemma 1 and we solve it via the virtual–surplus approach. The sufficient conditions in Lemma 1 are satisfied: marginal cost is constant, the revenue function is homothetic in output, and buyer types are Pareto–distributed. With one–dimensional type $\tau_{is'}$ and quasilinear transfers, expected revenue equals expected virtual surplus. Under the regularity condition $\rho^\ell_{ss'} > \sigma_{s'}$ (increasing virtual value), the problem separates across (ℓ, s) for a given seller sector s' and is solved pointwise in τ; transfers follow from the envelope formula with IR binding at $\underline{\tau}$.

Proposition 1 (Optimal Nonlinear Price for Upstream Sellers). *In equilibrium, the optimal contract offered by an upstream seller $j \in \mathcal{U}_{s'}$ to any buyer $i = (\ell, s, z_i)$ is a two–part tariff:*

$$T_{ij} \;=\; p^\ell_{js}\, m_{ij} \;+\; F^\ell_{js},$$

with marginal (allocative) price

$$p^\ell_{js} \;=\; \mu^\ell_{ss'}\, c_j, \qquad \mu^\ell_{ss'} \;=\; \frac{\rho^\ell_{ss'}}{\rho^\ell_{ss'} - 1}, \qquad \rho^\ell_{ss'} \;=\; \xi^\ell_s\, \sigma_{s'},$$

i.e., a constant markup over marginal cost within each buyer type–sector (ℓ, s) *for a given seller sector s'. The fixed component F^ℓ_{js} is chosen so that the lowest buyer type obtains zero surplus:*

$$F^\ell_{js} \;=\; \frac{1}{\sigma_{s'} - 1}\, \tau_{is'}(\underline{z}) \left(M_{is'}(\underline{z})\right)^{\frac{\sigma_{s'}-1}{\sigma_{s'}}} \left(\frac{p^\ell_{js}}{P^\ell_{ss'}}\right)^{1-\sigma_{s'}} \;=\; \left(\frac{z_j}{\widetilde{z}^u_{s'}}\right)^{\sigma_{s'}-1} \overline{F}^\ell_{ss'}, \qquad \tau_{is'} \;=\; P^\ell_{ss'}\, M^{1/\sigma_{s'}}_{is'}.$$

Here $\overline{F}^\ell_{ss'}$ denotes the average flat fee per seller *in sector s':*

$$\overline{F}^\ell_{ss'} \;\equiv\; \frac{1}{N^u_{s'}}\, \frac{1}{\sigma_{s'} - 1}\, \tau_{is'}(\underline{z}) \left(M_{is'}(\underline{z})\right)^{\frac{\sigma_{s'}-1}{\sigma_{s'}}}.$$

Proof Sketch and Verification of the Guesses. For upstream sellers (virtual surplus), fix a seller sector s' and a seller $j \in \mathcal{U}_{s'}$. Because the objective and constraints are additively separable across buyer partitions (ℓ, s), the mechanism is solved partition–by–partition. By Lemma 1 with type $\tau_{is'}$ (Pareto tail $\rho^\ell_{ss'}$), the partition problem is:

$$\max_{x(\tau)}\; N^\ell_s\, \mathbb{E}_{\tau_{is'}}\!\left[\left(\tau - g^{-1}(\tau)\right) \frac{\sigma_{s'}}{\sigma_{s'} - 1}\, x(\tau)^{\frac{\sigma_{s'}-1}{\sigma_{s'}}} \;-\; c_j\, x(\tau)\right].$$

For Pareto, $g^{-1}(\tau) = \tau/\rho^\ell_{ss'}$, so the virtual value is increasing when $\rho^\ell_{ss'} > \sigma_{s'}$. The FOC yields a constant markup within the partition:

$$p^\ell_{js} \;=\; \mu^\ell_{ss'}\, c_j, \qquad \mu^\ell_{ss'} \;=\; \frac{\rho^\ell_{ss'}}{\rho^\ell_{ss'} - 1},$$

and the fixed component is pinned down by IR at the lowest buyer type $\underline{z}^{\ell}_{s}$:

$$F^{\ell}_{js} \;=\; \frac{1}{\sigma_{s'}-1}\,\tau_{is'}(\underline{z}^{\ell}_{s})\left(M_{is'}(\underline{z}^{\ell}_{s})\right)^{\frac{\sigma_{s'}-1}{\sigma_{s'}}}\left(\frac{p^{\ell}_{js}}{P^{\ell}_{ss'}}\right)^{1-\sigma_{s'}}\;.$$

For upstream sellers (homothetic revenue), constant allocative prices $p^{\ell}_{js} = \mu^{\ell}_{ss'}c_j$ and the CES share rule imply, after aggregating over buyers and partitions, that total output Q_j is proportional to $c_j^{-\sigma_{s'}}$ times an aggregate demand term that depends on buyer masses, price indices, and materials bundles (see Appendix C.2). Equivalently, $c_j^{1-\sigma_{s'}}$ is proportional to $Q_j^{(\sigma_{s'}-1)/\sigma_{s'}}$. Both the per–unit revenue and the fee component scale with the same CES share, so total revenue scales as $R_j \;\propto\; Q_j^{(\sigma_{s'}-1)/\sigma_{s'}}$, up to a sector–$s'$ constant that aggregates buyer–side objects. This verifies Guess 1 (two–part tariffs) and Guess 2 (homothetic revenues) for upstream sellers.

For retailers ($\ell = r$), linear pricing under within–sector CES demand yields revenue proportional to $Q_j^{(\varphi_s-1)/\varphi_s}$ with a shifter depending on the retail price index P_s and expenditure $\theta_s Y$ (with $P_Y \equiv 1$), as shown in Appendix C.1. This completes the verification of Guess 2 for retailers and, together with the upstream case, confirms both guesses.

4.5 Other Pricing Regimes for Welfare Comparisons

We compare the nonlinear–pricing benchmark to two counterfactual policies: (i) monopolistic competition with uniform linear prices, which corresponds to a complete ban on price discrimination; and (ii) a planner–implemented allocation in a decentralized equilibrium, which is attained with a ban on price discrimination and an output subsidy that restore marginal cost pricing conditional on entry.

Monopolistic Competition (uniform linear pricing). Under uniform linear pricing, each upstream seller $j \in \mathcal{U}_{s'}$ charges the same CES markup over marginal cost to all buyers, regardless of their partition (ℓ, j):

$$p^{\ell,\mathrm{Lin}}_{js} \;=\; \mu^{\mathrm{Lin}}_{s'}\,c_j \qquad \mu^{\mathrm{Lin}}_{s'} \;\equiv\; \frac{\sigma_{s'}}{\sigma_{s'}-1},$$

where c_j is the marginal cost of seller firm j. Thus, in contrast to nonlinear pricing, allocative prices do not vary across buyer partitions but only by seller sector s'.

Retailers in sector s sell to final demand at the CES markup

$$\mu^{r,\mathrm{Lin}}_{s} \;=\; \frac{\varphi_s}{\varphi_s-1}.$$

For buyers of type ℓ in sector s, the CES price index for inputs from upstream sector q is

$$P_{ss'}^{\ell,\mathrm{Lin}} = \left(\int_{j \in \mathcal{U}_{s'}} \left(p_{js}^{\ell,\mathrm{Lin}} \right)^{1-\sigma_{s'}} \, d\nu_{s'}(j) \right)^{\frac{1}{1-\sigma_{s'}}} = \mu_{s'}^{\mathrm{Lin}} \left(\int_{j \in \mathcal{U}_{s'}} c_j^{1-\sigma_{s'}} \, d\nu_{s'}(j) \right)^{\frac{1}{1-\sigma_{s'}}},$$

where $d\nu_{s'}(j)$ integrates over active upstream firms in sector s', with free entry in each (ℓ, s).

Lemma 2 (Efficiency with CES markups and per–unit output subsidies)**.** *Consider the economy under a complete ban on price discrimination, so all sellers post uniform linear prices: each upstream seller sector $s' \in \mathcal{S}$ and retail sector $s \in \mathcal{S}$ sets the CES markup over marginal cost. The efficient allocation is achieved if the government rebates a per–unit output subsidy that restores marginal–cost pricing conditional on entry:*

$$p_j^{Lin} = \mu_{s'}^{Lin} c_j, \qquad \mu_{s'}^{Lin} = \frac{\sigma_{s'}}{\sigma_{s'} - 1}, \qquad \tau_{s'}^{u} = \left(1 - \frac{1}{\mu_{s'}^{Lin}} \right) c_j = \frac{1}{\sigma_{s'}} c_j;$$

$$p_i^{r,Lin} = \mu_s^{r,Lin} c_i, \qquad \mu_s^{r,Lin} = \frac{\varphi_s}{\varphi_s - 1}, \qquad \tau_s^{r} = \left(1 - \frac{1}{\mu_s^{r,Lin}} \right) c_i = \frac{1}{\varphi_s} c_i.$$

Then the resulting decentralized equilibrium is efficient.

The lemma is a special case of the general result in Theorem 1 of Baqaee and Farhi (2020) (details in Appendix C.3). Efficiency is obtained in a decentralized equilibrium when each variety charges a markup equal to its consumer–surplus ratio and receives output subsidies that exactly offset the induced within–period pricing wedge. In our CES setting, the consumer–surplus ratio for an upstream variety coincides $\mu_{s'}^{\mathrm{Lin}}$, and for a retail variety with $\mu_s^{r,\mathrm{Lin}}$. While charging the CES markup delivers the correct expected profits and thereby ensures efficient entry, it distorts input choices by acting as a tax on production. An output subsidy is therefore required to undo this distortion and restore marginal–cost pricing conditional on entry.

4.6 Theoretical Results

We now collect the main equilibrium implications of nonlinear pricing in our supply-chain model.

Result 1. Allocative Markups under Nonlinear Pricing. For any buyer partition (ℓ, s) and seller sector s', the allocative markup under nonlinear pricing is strictly below the linear–CES markup:

$$\mu_{ss'}^{\ell} = \frac{\rho_{ss'}^{\ell}}{\rho_{ss'}^{\ell} - 1} < \frac{\sigma_{s'}}{\sigma_{s'} - 1} \quad \text{since} \quad \rho_{ss'}^{\ell} = \xi_s^{\ell} \sigma_{s'} \text{ with } \xi_s^{\ell} > 1.$$

Under linear pricing, the markup is determined mechanically by the elasticity of substitution. With nonlinear pricing, it is instead governed by the inverse hazard rate of the buyer valuation distribution. This aligns marginal revenue more closely with the shape of demand, because price discrimination allows the seller to extract surplus through flat fees rather than distorting marginal allocations. Consequently, nonlinear pricing reduces allocative distortions at the margin relative to linear pricing.

Result 2. Seller–identity invariance of the total unit markup. Fix a seller sector s' and a buyer partition (ℓ, s). For any seller $j \in \mathcal{U}_{s'}$ and buyer $i = (\ell, s, z_i)$, the per–unit payment decomposes as:

$$\frac{T_{ij}}{m_{ij}} = p^{\ell}_{js} + \frac{F^{\ell}_{js}}{m_{ij}}.$$

The resulting total unit markup (unit price over marginal cost) satisfies:

$$\frac{\frac{T_{ij}}{m_{ij}}}{c_j} = \mu^{\ell}_{ss'}\Big(1 + \chi^{\ell}_{ss'}(i)\Big), \qquad \mu^{\ell}_{ss'} = \frac{\rho^{\ell}_{ss'}}{\rho^{\ell}_{ss'} - 1},$$

where $\chi^{\ell}_{ss'}(i)$ is a buyer–specific scalar defined in Appendix C.4. It depends on (ℓ, s) objects (the sectoral price index and the buyer's sector–s' bundle, including the lowest buyer type) but not on the seller j. Hence, within a given buyer partition, the total unit markup is invariant to the seller's identity.

Result 3. Average flat fee paid to seller sector s' determinants For any buyer partition (ℓ, s) with lowest type $\underline{z}^{\ell}_s$, the average flat fee paid to seller sector s' is

$$\overline{F}^{\ell}_{ss'} = \frac{\psi^{\ell}_s}{\sigma_{s'} - 1}(1 - \alpha^{\ell}_s)\,\theta^{\ell}_{ss'}\,\frac{R^{\ell}_s(\underline{z}^{\ell}_s)}{N^{u}_{s'}} = \frac{P^{\ell}_{ss'}\,M_{is'}(\underline{z}^{\ell}_s)}{N^{u}_{s'}\,(\sigma_{s'} - 1)}.$$

Here $R^{\ell}_s(\underline{z}^{\ell}_s)$ denotes the revenue of the lowest–type buyer in (ℓ, s), $M_{is'}(\underline{z}^{\ell}_s)$ the corresponding sector–s' materials bundle, $P^{\ell}_{ss'}$ the sectoral price index faced by (ℓ, s), and $N^{u}_{s'}$ the mass of active upstream sellers in s'.

Five forces shape flat fees from buyers in (ℓ, s) to seller sector s': (i) lowest–type revenue $R^{\ell}_s(\underline{z}^{\ell}_s)$ raises extractable rents; (ii) revenue curvature ψ^{ℓ}_s (from $R_i = A^{\ell}_s Q_i^{\psi^{\ell}_s}$) scales marginal surplus, more concavity lowers fees; (iii) input importance $(1 - \alpha^{\ell}_s)\theta^{\ell}_{ss'}$ increases the surplus a seller can extract; (iv) a larger $\sigma_{s'}$ (greater substitutability) reduces $\frac{1}{\sigma_{s'}-1}$ and thus fees; (v) a larger $N^{u}_{s'}$ dilutes rents across more sellers, lowering the average fee.

Result 4. Firm Profits Rely on Flat Fees. With two–part tariffs, profits decompose cleanly into a marginal (per–unit) component and a fixed (flat–fee) component. This holds for upstream sellers and for retailers. The decomposition highlight which forces move profits: allocative markups on the margin, and the incidence of fixed transfers across buyer–seller pairs. For an upstream seller

$j \in \mathcal{U}_{s'}$:

$$\mathbb{E}\left[\Pi_j^u\right] = \underbrace{\sum_{\ell\in\{u,r\}}\sum_{s\in\mathcal{S}}\int_{i\in\mathcal{F}_{\ell s}}\left(p_{js}^{\ell}-c_j\right)m_{ij}\,d\nu_{\ell s}(i)}_{\text{allocative margin}} + \underbrace{\sum_{\ell\in\{u,r\}}\sum_{s\in\mathcal{S}}\int_{i\in\mathcal{F}_{\ell s}} F_{js}^{\ell}(i)\,d\nu_{\ell s}(i)}_{\text{flat–fee revenue}} - \underbrace{\sum_{t\in\mathcal{S}}\int_{h\in\mathcal{U}_t} F_{h\,s'}^{u}(j)\,d\nu_t(h)}_{\text{flat–fee payments}}.$$

For a retailer $i \in \mathcal{R}_s$:

$$\mathbb{E}\left[\Pi_i^r\right] = \underbrace{\left(\frac{1}{\varphi_s}\right)R_i}_{\text{allocative margin}} - \underbrace{\sum_{s'\in\mathcal{S}}\int_{j\in\mathcal{U}_{s'}} F_{js}^{r}(i)\,d\nu_{s'}(j)}_{\text{fees to upstream}}.$$

Realized profits depend on the specific network of trading partners (which sellers a buyer contracts with, and which upstream tiers a seller sources from). For sector–level analysis we therefore work with expected profits—i.e., averages over the equilibrium measures $\nu_{\ell s}$ (buyers in partition (ℓ, s)) and $\nu_{s'}$ (sellers in sector s'), which collapse partner–specific details into sectoral aggregates.

Here p_{js}^{ℓ} is the allocative (marginal) price charged by seller j to buyers in (ℓ, s), c_j is seller j's marginal cost, m_{ij} is buyer i's quantity purchased from j, and $F_{js}^{\ell}(i)$ is the flat fee paid by buyer i to seller j. The set $\mathcal{F}_{\ell s}$ collects active buyers in (ℓ, s) with measure $\nu_{\ell s}$; $\mathcal{U}_t$ is the set of upstream sellers in sector t with measure ν_t. The payment $F_{h\,s'}^{u}(j)$ denotes the flat fee that seller j (as a buyer of type u in buyer sector s') pays to its upstream supplier $h \in \mathcal{U}_t$.

The allocative margin captures the usual markup–cost wedge times purchased quantities; under nonlinear pricing, Result 1 implies these markups are lower than under uniform linear pricing, shrinking this component. The flat–fee terms redistribute surplus based on each seller's CES share in a buyer's materials bundle and on how important inputs are in production (via the θ's). In expectation, the network of bilateral contracts integrates out to sectoral objects, so expected upstream profits can be expressed as affine functions of sectoral labor expenditures (Appendix C.5), and profits of lowest–productivity retailers hinge on input substitutability and input cost shares (Appendix C.5). allocative markups.

The fact that profits need not vanish contrasts with the standard mechanism-design benchmark, in which the lowest type's surplus is pinned to zero. Here, bilateral surplus is zero at the margin of each input transaction, but integrating across all transactions leaves residual profits (through labor, which is not price-discriminated) or, conversely, negative profits when intermediates are insufficiently substitutable. In the limit, each supplier's marginal contribution equals the buyer's total surplus, so every supplier attempts to appropriate the full rent. This drives the lowest type's profit below zero and generates a hold-up problem that deters entry.

Flat fees are infra-marginal: they do not affect first-order input choices or final demand, but they reallocate surplus along the chain. With a representative owner, these transfers net out at a point in time; in general equilibrium, however, they shift free-entry conditions across (ℓ, s) altering the mass of active varieties and sectoral price indices. Hence welfare in the counterfactuals will be shaped by two channels: (i) changes in allocative markups (marginal wedges), and (ii) the reallocation of flat-fee income that tilts entry across sectors. The next section formalizes these channels and maps welfare changes to sufficient statistics tied to markups, price indices, and entry margins.

4.7 Welfare Decomposition: Intensive vs. Extensive Margins

We measure welfare by the inverse final price index,

$$W \equiv \frac{1}{P_Y}, \qquad \log P_Y = \sum_{s \in \mathcal{S}} \theta_s \log P_s.$$

With wage normalization and free entry, the representative household's income equals the wage, so $W = 1/P_Y$ (see Appendix C.6 for details). Within each retail sector s, the sectoral price index satisfies

$$P_s = \mu_s^r \, \Theta_s^r \, w^{\alpha_s^r} \left(\prod_{s' \in \mathcal{S}} \left(P_{s's}^r \right)^{(1-\alpha_s^r)\, \theta_{ss'}^r} \right) \left(N_s^r \right)^{-\frac{1}{\varphi_s - 1}} \mathcal{V}_s, \qquad P_{s's}^r = \mu_{s's}^r \, C_{s'}, \tag{15}$$

where μ_s^r is the retail-to-consumer markup (allocative wedge) in sector s, $P_{s's}^r$ is the marginal price paid by retail sector s to upstream sector s', $\mu_{s's}^r$ is the buyer-specific markup charged by s' to s, $C_{s'}$ is the upstream sector-s' marginal cost index, N_s^r is the mass of active retail varieties in (r, s), and $\mathcal{V}_s$ is the CES selection term capturing composition effects among active varieties.

For each upstream seller sector s', marginal cost satisfies the CES–CD recursion

$$C_{s'} = \Theta_{s'}^u \, w^{\alpha_{s'}^u} \left(\prod_{v \in \mathcal{S}} \left(P_{vs'}^u \right)^{(1-\alpha_{s'}^u)\, \theta_{s'v}^u} \right) \left(N_{s'}^u \right)^{-\frac{1}{\sigma_{s'} - 1}} \mathcal{V}_{s'}^u, \qquad P_{vs'}^u = \mu_{vs'}^u \, C_v, \tag{16}$$

where $P_{vs'}^u$ is the marginal price paid by upstream sector s' to upstream sector v, $\mu_{vs'}^u$ is the buyer-specific markup charged by v to s', $N_{s'}^u$ is the mass of active upstream varieties in (u, s'), $\sigma_{s'} > 1$ is the upstream CES elasticity, and $\mathcal{V}_{s'}^u$ is the upstream CES selection term.

Taking logs of (16) and substituting $P_{vs'}^u = \mu_{vs'}^u C_v$ gives

$$\log C_{s'} = \sum_{v \in \mathcal{S}} (1 - \alpha_{s'}^u)\, \theta_{s'v}^u \left(\log \mu_{vs'}^u + \log C_v \right) + \alpha_{s'}^u \log w + \log \Theta_{s'}^u - \frac{1}{\sigma_{s'} - 1} \log N_{s'}^u + \log \mathcal{V}_{s'}^u. \tag{17}$$

Stacking (17) across upstream sectors and collecting the input-cost shares into A^{uu} with entries $A_{s'v}^{uu} := (1 - \alpha_{s'}^u)\, \theta_{s'v}^u$ yields the fixed-point system

$$\log C^u = A^{uu} \log C^u + \log \mu^{uu} + \alpha^u \log w + \log \Theta^u - \frac{\log N^u}{\sigma - 1} + \log \mathcal{V}^u, \tag{18}$$

where $\log \mu^{uu}$ stacks the upstream to upstream buyer-specific wedges, division by $(\sigma-1)$ is elementwise, and vectors are conformable by sector. In changes across regimes, constants (technology) and the wage numeraire drop out, and the upstream cost response is

$$\Delta \log C^u \;=\; \Big(I - A^{uu}\Big)^{-1}\!\left(\Delta \log \mu^{uu} \;-\; \frac{\Delta \log N^u}{\sigma - 1} \;+\; \Delta \log \mathcal{V}^u\right). \tag{19}$$

Substituting (19) into (15) below will deliver the welfare decomposition by loading upstream terms first at the retail interface and then through the upstream network.

Input–output objects and final-demand exposures. Define the matrix A^{uu} as the collection of upstream–upstream cost shares with entries $A^{uu}_{s'v} := (1 - \alpha^u_{s'})\,\theta^u_{s'v}$ and let the matrix B^{ru} to collect retail–upstream cost shares with entries $B^{ru}_{ss'} := (1 - \alpha^r_s)\,\theta^r_{ss'}$.

The vector $\tilde{\lambda}_{ru}$ gives, component by component, the share of final demand that is spent—through retailers—on each upstream sector's composite; it measures how a one-percent change in an upstream sector's marginal cost, holding upstream propagation fixed, loads into the final price index at the retail interface. The vector $\tilde{\lambda}_u$ applies the Leontief inverse to $\tilde{\lambda}_{ru}$ and therefore adds all direct and indirect upstream linkages; each component is the cost-based exposure of final demand to that upstream sector after accounting for the entire upstream network. These exposures are the exact multipliers that map upstream markups, variety, and selection into welfare in Proposition 2 (see Appendix C.6 for detail derivation.

$$\tilde{\lambda}_{ru} := \theta^\top B^{ru} \in \mathbb{R}^{1\times|\mathcal{S}|}, \qquad \tilde{\lambda}_u := \tilde{\lambda}_{ru}\,(I - A^{uu})^{-1} \in \mathbb{R}^{1\times|\mathcal{S}|}. \tag{20}$$

Proposition 2 (Exact welfare decomposition)**.** *Starting from* (15) *with $P^r_{s's} = \mu^r_{s's} C_{s'}$ and the upstream marginal-cost recursion, the change in welfare satisfies the exact identity:*[19]*.*

$$W = \underbrace{-\sum_{s\in\mathcal{S}} \theta_s\,\Delta \log \mu^r_s \;-\; \tilde{\lambda}_{ru}\,\Delta \log \mu^r \;-\; \tilde{\lambda}_u\,\Delta \log \mu^{uu}}_{\textit{Intensive (allocative) markups: retail–consumer, retail–upstream, upstream–upstream}}$$

$$+ \underbrace{\sum_{s\in\mathcal{S}} \frac{\theta_s}{\varphi_s - 1}\,\Delta \log N^r_s \;+\; \tilde{\lambda}_u\left(\frac{\Delta \log N^u}{\sigma - 1}\right)}_{\textit{Extensive (variety/masses)}} - \underbrace{\sum_{s\in\mathcal{S}} \theta_s\,\Delta \log \mathcal{V}_s \;-\; \tilde{\lambda}_u\,\Delta \log \mathcal{V}^u}_{\textit{Selection (composition)}}.$$

The first line captures the allocative effect of markups; the second collects the variety effect at retail and upstream; the last line captures selection (productivity composition among active varieties). When the

[19] Dimensions: $W \in \mathbb{R}$ is a scalar. $\theta \in \mathbb{R}^{|\mathcal{S}|}$ (used as $\theta^\top$) collects final-expenditure shares. $A^{uu}, B^{ru} \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{S}|}$ are cost-share matrices. $\tilde{\lambda}_{ru}, \tilde{\lambda}_u \in \mathbb{R}^{1\times|\mathcal{S}|}$ are row vectors (exposures). $\mu^r, \mu^{uu}, N^r, N^u, \mathcal{V}, \mathcal{V}^u \in \mathbb{R}^{|\mathcal{S}|}$ are sector-level vectors with components $\mu^r_s, \mu^{uu}_{s'}, N^r_s, N^u_{s'}, \mathcal{V}_s, \mathcal{V}^u_{s'}$. Elasticities $\varphi = (\varphi_s)_s$ and $\sigma = (\sigma_{s'})_{s'}$ are vectors; division by $(\sigma - 1)$ below is elementwise. All $\Delta \log(\cdot)$ operations act componentwise on vectors

composition is invariant across the comparison (for example, Pareto-tail invariance), the selection terms are zero.

The decomposition isolates three forces. First, allocative: lower markups improve welfare, with impact scaled by consumer shares, the retail–upstream interface exposure, and the full upstream exposure. Second, extensive: larger masses of active firms raise welfare, with strength governed by the same exposures. Third, selection: changes in the productivity composition of active varieties also load through these exposures and disappear when composition is unchanged. The exposure vectors act as cost-based Domar multipliers, translating sector-level shocks into aggregate consequences and serving as practical sufficient statistics for counterfactual analysis.

Welfare ratios across price regimes. Using Proposition 2, we compare nonlinear pricing (NLP) to uniform linear pricing (Lin) along the supply chain. Assume retail–consumer markups coincide across regimes and that productivities are Pareto so all types are served; selection terms then drop. Let $\mu^{r,R} \in \mathbb{R}^{|\mathcal{S}_u|}$ and $\mu^{uu,R} \in \mathbb{R}^{|\mathcal{S}_u|}$ denote the sector-level upstream–retail and upstream–upstream wedge vectors under regime $R \in \{\text{NLP}, \text{Lin}\}$, and let $N_s^{r,R}$ and N_u^R be the retail and upstream masses. Then:

$$\frac{W^{\text{NLP}}}{W^{\text{Lin}}} = \underbrace{\prod_{u\in\mathcal{S}_u}\left(\frac{\mu_u^{r,\text{NLP}}}{\mu_u^{r,\text{Lin}}}\right)^{-\tilde{\lambda}_{ru}} \prod_{u\in\mathcal{S}_u}\left(\frac{\mu_u^{uu,\text{NLP}}}{\mu_u^{uu,\text{Lin}}}\right)^{-\tilde{\lambda}_u}}_{\text{Intensive factor}} \underbrace{\prod_{s\in\mathcal{S}_r}\left(\frac{N_s^{r,\text{NLP}}}{N_s^{r,\text{Lin}}}\right)^{\frac{\theta_s}{\varphi_s-1}} \prod_{u\in\mathcal{S}_u}\left(\frac{N_u^{\text{NLP}}}{N_u^{\text{Lin}}}\right)^{-\frac{\tilde{\lambda}_u}{1-\sigma_u}}}_{\text{Extensive factor}} .$$

Nonlinear pricing weakens allocative wedges relative to linear pricing (two–part tariffs raise buyer elasticities and lower markups), so the intensive factor exceeds one, with gains scaled by the final-demand exposures $\tilde{\lambda}_{ru}$ and $\tilde{\lambda}_u$. Flat fees are infra-marginal and operate only through participation and entry; the extensive factor can therefore amplify or offset the intensive gains depending on how firm masses adjust in highly exposed sectors. The net effect is thus disciplined by the exposure maps and general-equilibrium entry responses, which we quantify next.

5 Model Calibration and Quantification

5.1 Parameters Estimation

We use Chilean administrative microdata (2005–2022): firm accounts (revenues, wage bill, headcounts, profits, capital) and the universe of firm-to-firm transactions (quantities, prices, counterparties, locations). Parameters tied to technology or marginal costs are measured at fine granularity (6-digit sector × firm type) and mapped to the model's 11-sector blocks when needed. To ensure observed unit prices are allocative (rather than flat-fee affected), most moments are computed on large firms and then averaged over time. Table 3 shows the seven parameters we estimate, the method used, and their granularity. Appendix D provides detailed explanations on estimation.

Table 3: Estimated Parameters

Parameter	Strategy	Granularity
Labor output elasticity (α_s)	Calibrated from data	626 sectors × firm type
Final demand elasticity (θ_r)	Calibrated from data	626 sectors
Input-Output elasticity (θ_{iu})	Calibrated from data	626 sectors × firm type
Final demand bundle elasticity (φ)	Pin down by CES results and data	11 sectors
Material bundle elasticity (σ)	Covid shock for Chile estimation	11 sectors
Exit rate(δ)	Calibrated from data	626 sectors
Entry cost (c_e)	Pin down by free entry and data	626 sectors × firm type
Productivity Pareto tail (κ)	MLE estimation	11 sectors × firm type

Labor output elasticity α_s. This parameter is the Cobb–Douglas weight on primary inputs (labor plus the user cost of capital) in production. We recover it from firm accounts as the non-material cost share at the 6-digit sector and firm-type level, restricting to large firms to align observed unit prices with marginal prices and winsorizing extremes for stability. Because flat fees are small for these firms, variable-cost shares are reliable proxies for total-cost shares. α_s governs how sectoral output responds to wages relative to materials prices: higher α_s amplifies labor-market shocks and dampens pass-through from input-price shocks.

Input-Output elasticity θ_r. These Cobb–Douglas weights allocate the representative consumer's expenditure across retail sectors. With linear pricing to final consumers, retailer revenues identify sectoral expenditure; we form each sector's share of aggregate retail sales using large firms and average across years. These shares anchor the final-demand system and the welfare accounting used in counterfactuals.

Input-Output elasticity θ_{iu}. These are buyer-facing expenditure shares on upstream seller sectors within the materials bundle. Using transaction-level data, we compute for each buyer the fraction of variable materials spending sourced from each upstream sector, aggregate to 6-digit industries within year, and average over time. The resulting matrix provides the micro foundation of the input–output network, determining exposure patterns and the scope for intensive-margin substitution when relative prices move.

Material bundle elasticity σ. This elasticity measures how easily buyers substitute across varieties within an upstream seller sector in response to relative price changes. We exploit the quasi-experimental disruptions from Chile's early COVID-19 lockdowns (March 2020) by instrumenting the main pre-shock supplier's relative price change with that supplier's lockdown exposure, estimating sector-specific elasticities via two-stage least squares on 12-month differences, focusing on large buyers, and excluding cases with potentially confounded exposure (buyer location, buyer customers, or other inputs locked down).[20] A higher σ implies rapid rewiring and strong intensive-margin reallocation with muted pass-through; a lower σ implies stickier relationships, larger infra-marginal surplus, and greater scope for rent extraction via flat fees.

Final demand bundle elasticity φ. This is the elasticity of substitution across retail varieties within a sector and, under linear pricing, coincides with the inverse markup wedge. We recover it from sectoral accounts implied by CES demand, linking pooled sectoral sums of revenues, profits, and fixed operating costs (in labor units) for large retailers and averaging across years. Higher φ indicates keener competition and smaller allocative wedges; lower φ sustains higher markups and larger deadweight losses for a given cost shock.

Exit rate δ. The exit rate is the one-year hazard that an active firm ceases operations. We compute it at the 6-digit sector interacted with firm-type level by tracking the share of firms present in a given year that are not observed in the following year, and then averaging over 2005–2022. This object disciplines the expected lifespan of an entrant and, together with the discount rate, determines how quickly future profits are attenuated. Higher δ raises the payoff required to justify entry, thins steady-state firm mass for given fundamentals, and shifts the balance between churn and scale. In counterfactuals, sectors with elevated hazards display weaker persistence of shocks and a larger role for extensive-margin adjustments.

Entry cost c_e. Entry costs are the labor-measured sunk resources required to create an operating firm. We combine sector–type averages of accounting profits and wages with the empirically estimated exit hazards and a standard discount rate to obtain the expected present value of a surviving firm; free entry equates that value, scaled by the share of positive-profit firms, to the labor cost of entry. We report both currency units and "wage-bill equivalents" for comparability across sectors. Higher c_e depresses equilibrium firm mass and raises average scale, sharpening how nonlinear pricing interacts with the extensive margin and rent allocation across links.

Productivity Pareto tail κ. This exponent captures the thickness of the upper tail of firm productivity. We estimate Pareto tail exponents for firm employment by maximum likelihood in the upper tail and map them to productivity tails using the model's monotone link between productivity and employment, whose slope is governed by σ. Thinner tails (larger κ) limit selection

[20] For sectors with estimates below one, we conservatively adopt the smallest credible value above one.

and reallocation gains; fatter tails (smaller κ) make dispersion central for welfare and shape how nonlinear pricing shifts surplus across the distribution.

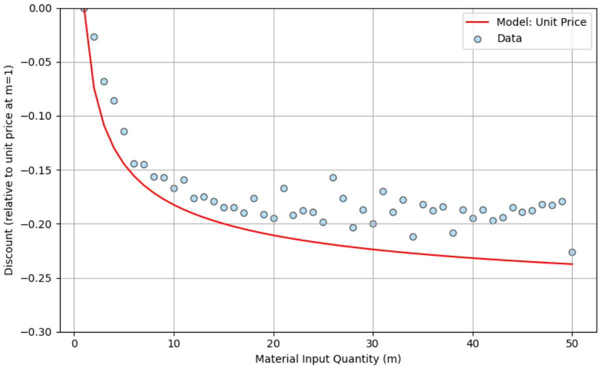
Together, these objects pin down (i) how easily buyers and consumers reallocate across varieties (σ, φ); (ii) how strongly sectors load on wages versus materials costs (α); (iii) who buys what from whom and in what proportion (θ); (iv) how many firms enter and survive (c_e and δ); and (v) how dispersed productivity is within sectors (κ). This configuration determines the balance between intensive reallocation and extensive entry/exit in counterfactuals and the extent to which nonlinear pricing redistributes surplus without distorting marginal allocations.

5.2 Model Quantification

Model Fit. Once calibrated, we assess the model's empirical validity by comparing the equilibrium nonlinear unit prices to those observed in the data. Figure 5 contrasts model-implied quantity discounts with the empirical strategy from the descriptive evidence section. The figure plots unit prices charged by the average upstream firm to retailers as a function of purchased quantity, mapping to productivity types z from a low reference z_0 up to the 99th percentile of the z distribution, with quantities normalized to the interval $[1, 50]$.

The nonlinear pricing pattern in the data is not targeted in calibration; parameters are estimated independently. Nevertheless, the model reproduces the negative unit price–quantity gradient: larger purchases receive larger discounts. It also matches discount magnitudes, which reach roughly 24% for large purchases in both model and data.

Figure 5: Calibrated Model Unit Prices vs. Data Fixed-Effects Regression



Notes: The figure compares model-predicted quantity discounts with observed patterns in the data. Model unit discounts are computed for the average upstream price schedule to retailers, normalizing continuous input quantity to range from 1 to 50. The model reproduces both the shape and magnitude of discounts without explicitly targeting these moments, reaching approximately 24% for large purchases in both model and data.

Aggregate Welfare Ratios. We compare welfare across pricing regimes relative to the efficient benchmark. Across regimes we keep CES markups from retailers to final consumers fixed and allow upstream markups (to retailers and to other upstream firms) to vary; all endogenous objects are recomputed in separate equilibria. With Pareto distributed firm types, every productivity type is served under each regime; this shuts down selection from Proposition 2 and makes the composition of active firms invariant. Under wage normalization $w \equiv 1$, welfare equals the inverse of the final price index, $\mathcal{W} = 1/P_Y$.

Table 4: Aggregate welfare (relative to efficiency)

Price Regime	$\mathcal{W}^{R}/\mathcal{W}^{\text{Eff}}$	$\mathcal{W}^{\text{NLP}}/\mathcal{W}^{\text{Lin}}$
Nonlinear (NLP)	0.745	1.534
Linear pricing (Lin)	0.486	

Notes: Welfare is computed as $1/P_Y$ with $w \equiv 1$. Retail→consumer markups are held fixed; upstream markups differ by regime: NLP uses two–part tariffs with buyer-specific μ, Lin uses uniform CES $\mu = \sigma/(\sigma - 1)$, and Eff implements marginal-cost pricing. Entry and firm masses are re-solved in each regime.

Nonlinear pricing achieves approximately 75% of efficient welfare, while linear pricing achieves approximately 50% . Moving from linear to nonlinear pricing thus closes half of the efficiency gap

Free entry responds to regime-specific profitability through markups and flat fees, determining steady-state firm masses by sector. These masses enter CES price indices via variety effects: higher masses reduce price indices and raise welfare. Flat fees affect welfare only through the extensive margin by shifting expected profits and firm entry. With selection fixed, the welfare gap relative to efficient pricing decomposes into an intensive component (markup accumulation along the supply chain) and an extensive component (entry-driven variety), each weighted by final-demand exposure that propagates effects through the supply chain.

Aggregate intensive vs. extensive contributions and entry responses. We quantify how pricing regimes differ from efficient pricing through two channels. The intensive channel summarizes allocative wedges from markups propagated by final demand exposure maps along the supply chain; the extensive channel summarizes general–equilibrium entry responses and the implied masses of active firms. The factors in Table 5 are computed from Proposition 2 and shares are computed on absolute log contributions because intensive and extensive forces can move in opposite directions.[21]

[21] We proportionally adjust the factors so that their product reported in Table 4 exactly matches the welfare ratio relative to the efficiency in each regime, the residual between the product and the reported welfare ratio is ¡0.03 in levels.

Table 5: Aggregate decomposition and firm masses (relative to efficiency)

Price Regime	Intensive	Extensive	Share$_{\text{int}}$	Share$_{\text{ext}}$	Upstream mass	Retail mass
Nonlinear	0.67	1.12	0.79	0.21	1.18	1.17
Linear	0.46	1.06	0.93	0.07	1.00	1.44

Notes: Intensive and Extensive are calibrated factors consistent with the welfare ratio $\mathcal{W}^R/\mathcal{W}^{\text{Eff}}$; the residual between the product of adjusted factors and the reported welfare ratio is below 0.03 in levels. Shares are fractions of absolute log contributions attributed to intensive vs. extensive components. Mass ratios are $N^{u,R}/N^{u,\text{Eff}}$ and $N^{r,R}/N^{r,\text{Eff}}$. All entries rounded to two decimals.

Two patterns stand out. Intensive distortions dominate the welfare losses: under nonlinear pricing about 79% of the absolute deviation is intensive, and under linear pricing about 93%. The extensive margin is pro–competitive in both regimes (factors above one), but quantitatively modest relative to the intensive losses, especially under linear pricing. On the entry side, nonlinear pricing raises firm masses in both firm types by roughly 17% relative to efficiency, whereas linear pricing leaves upstream nearly unchanged and expands retail sharply. Despite this retail expansion, linear–pricing welfare remains below nonlinear because uniform linear markups create a stronger allocative wedge that dominates the overall outcome.

Higher markups raise expected profits and stimulate entry, but entry absorbs labor as a fixed cost; equilibrium adjusts toward more firms producing less, so the CES variety force lowers the price index only partially and cannot offset markup accumulation along the supply chain. The dominance of intensive losses is amplified in sectors with large final consumption exposure weights, where upstream markups load more heavily into final prices.

Nonlinear vs. Linear Pricing: Opening Welfare Ratios by Sector. The sectoral opening is governed by how strongly sectors are exposed to final demand through the supply chain. We use two sufficient statistics to capture final demand exposure of sectors, $\tilde{\lambda}_{ru}$ summarizes the direct and indirect exposure of the retail sector r into final demand through upstream sector u, and $\tilde{\lambda}_u$ account for the direct and indirect exposure of upstream sector u into final demand. Their construction and levels are reported in Appendix E.1. In our data, final demand exposure is concentrated in a few sectors: Retail & Wholesale, Manufacturing, Transport/ICTs, and Construction (see Appendix E.1), so compressing markups in these sectors delivers disproportional aggregate gains. The sectoral results below mirror this pattern: the intensive component of nonlinear pricing loads on sectors more exposed to final demand, while extensive responses are smaller and mixed across sectors.

Taking logs of the welfare–ratio identity of nonlinear prices to linear prices yields the sector–additive decomposition used in the welfare section:

$$\log \frac{\mathcal{W}^{\mathrm{NLP}}}{\mathcal{W}^{\mathrm{Lin}}} = \underbrace{\sum_u \Big[-\tilde{\lambda}_{ru}\, \Delta \log \mu^{r}_{u} \Big]}_{\text{Intensive: retailers}} + \underbrace{\sum_u \Big[-\tilde{\lambda}_{u}\, \Delta \log \mu^{uu}_{u} \Big]}_{\text{Intensive: upstream}} + \underbrace{\sum_s \frac{\theta_s}{\varphi_s - 1}\, \Delta \log N^{r}_{s}}_{\text{Extensive: retailers}} + \underbrace{\sum_u \frac{\tilde{\lambda}_u}{\sigma_u - 1}\, \Delta \log N^{u}_{u}}_{\text{Extensive: upstream}} .$$

The first two sums are the upstream intensive contributions, splitting allocative markups into upstream–retail and upstream–upstream components; the last two sums are the extensive margin contributions from retail and upstream masses. Positive entries indicate that nonlinear pricing raises welfare relative to linear pricing in that sector through lower markups (intensive margin) or different firm variety masses (extensive margin).

Table 6 shows that nonlinear prices unambiguously improve welfare across all sectors relative to linear prices on the intensive margin due to attenuated double marginalization. On the extensive margin among upstream firms, nonlinear prices are generated, except for three sectors, in larger firm masses. In contrast, there is unambiguously more entry in linear pricing on the retailers' extensive margins. This is consistent with flat fee distortions, which are collected and paid by upstream firms and only paid by retailers in nonlinear setups.

Allocative improvements in the intensive margin are split between within-upstream links (I^{uu}) and upstream to retail links (I^{ur}) . Within upstream links, Retail & Wholesale contributes 0.068 (~ 38% of the column total 0.181), followed by Manufacturing (16%), Transport & ICTs (13%), and Construction (12%). On upstream to retail links, the leaders are Construction (33%) , Retail & Wholesale (20%), and Manufacturing (13%). Nonlinear prices lower marginal prices bite hardest where intermediate intensity and exposure to downstream demand are large.

The extensive margin is net positive but smaller: upstream variety rises while retail variety falls in nonlinear prices relative to linear prices, yielding a net of 0.049. Upstream expansion is concentrated in Construction (60% of upstream entry contribution, Utilities (17%), Real Estate Services (12%), and Mining (7%). Retail contraction is concentrated in Financial Services (42% of retail entry contribution), Construction (14%), and Transport & ICTs (14%). This is consistent with flat fee-based rent extraction discouraging retail entry, while improved margins and fee income shift profits and entry upstream.

Netting all four columns, the main sectors explaining welfare difference in favor of nonlinear prices relative to linear prices are Construction (0.173) and Retail & Wholesale (0.101), followed by Utilities (0.050) and Manufacturing (0.046). Financial Services (−0.041) and Business Services (−0.001) are the only net sectors where linear prices improve welfare relative to nonlinear prices; elsewhere, the improvement is near-zero or positive.

Table 6: Nonlinear Prices Relative to Linear, by Sector and Margin (log contributions)

Sector	Intensive (allocative)		Extensive (variety)		net NLP/Lin
	$\log I^{uu}$	$\log I^{ur}$	$\log E^{u}$	$\log E^{r}$	
Agriculture	0.010	0.010	0.005	-0.003	0.022
Mining	0.003	0.003	0.014	-0.001	0.019
Manufacturing	0.029	0.024	0.002	-0.009	0.046
Utilities	0.006	0.016	0.032	-0.004	0.050
Construction	0.022	0.059	0.112	-0.020	0.173
Retail and Wholesale	0.068	0.036	0.005	-0.008	0.101
Transport and ICTs	0.023	0.007	0.000	-0.019	0.011
Financial Services	0.008	0.012	-0.002	-0.059	-0.041
Real Estate Services	0.004	0.009	0.023	-0.004	0.032
Business Services	0.006	0.005	-0.001	-0.011	-0.001
Personal Services	0.001	0.001	-0.000	-0.002	0.000
Sum	0.181	0.180	0.188	-0.139	0.410

Notes: Positive log entries mean NLP raises welfare relative to Lin through that sector–margin. $\log I^{uu}$ and $\log I^{ur}$ are the upstream–upstream and upstream–retail allocative components; the aggregate intensive log equals the sum of their column totals. $\log E^{u}$ and $\log E^{r}$ are exposure–weighted variety components from upstream and retail masses; their column totals add to the aggregate extensive log. Retail–consumer markups are identical across regimes, so retail intensive terms are zero. Column and row sums match the aggregate intensive and extensive logs reported in Table 5(up to rounding)

Three messages stand out for our quantitative exercise. First, nonlinear pricing delivers most of its welfare advantage through intensive relief of wedges at the most exposed upstream sectors; this component explains the bulk of the aggregate gap between regimes. Second, extensive forces are reinforcing but modest: entry expands upstream and contracts at retail, yet these effects are small relative to the intensive gains. Third, the exposure to final demand statistics provides actionable guidance: they point to where compressing markups and shifting participation margins move the aggregate needle. Taken together with the efficient benchmarks, these results show that nonlinear prices can recoup a large share of the efficiency loss from uniform linear markups, especially when applied in highly exposed to final demand upstream sectors.

6 Conclusion

Using universe-level firm-to-firm transactions from Chile, we reject linear pricing and document hybrid price discrimination combining second-degree (quantity discounts) with third-degree (buyer group) elements. The pricing patterns are isomorphic with two-part tariffs: common marginal prices within seller-product pairs combined with buyer-specific flat fees. This hybrid structure appears throughout the supply chain, with quantity driving price curvature while buyer characteristics shift levels and slopes

In our general equilibrium framework, nonlinear pricing improves allocations relative to linear pricing by lowering marginal prices, while flat fees redistribute rents and distort entry. Our welfare decomposition shows the net effect is positive: nonlinear pricing achieves 75 percent of efficient welfare versus 50 percent under linear pricing. The main takeaway is that average prices, if not allocative, can mislead welfare assessments. What matters are marginal allocative prices and the extent of rent extraction across the supply chain.

These results imply a policy design principle: do not uniformly ban quantity discounts. Instead, target markup accumulation in supply chains and limit rent extraction to preserve low marginal prices without discouraging entry. Enforcement should focus on sectors where marginal-average price wedges are large and final demand exposure is high, rather than on average price differentials.

Our framework provides a practical methodology for policy evaluation using standard transaction-level data. It enables ex-ante assessment of quantity discount regulations and ex-post evaluation of sector-specific contracting rules through welfare-sufficient statistics that separate allocative from variety effects. While our quantitative findings depend on Chilean market structure, the approach applies wherever similar microdata exists, offering an empirical foundation for price discrimination policy.

References

Acemoglu, D., Carvalho, V. M., Ozdaglar, A., and Tahbaz-Salehi, A. (2012). The network origins of aggregate fluctuations. *Econometrica*, 80(5):1977–2016.

Baqaee, D. and Farhi, E. (2020). Entry vs. rents: Aggregation with economies of scale. Technical report, National Bureau of Economic Research.

Boehm, J., South, R., Oberfield, E., and Waseem, M. (2024). The network origins of firm dynamics: Contracting frictions and dynamism with long-term relationships.

Bornstein, G. and Peter, A. (2024). Nonlinear pricing and misallocation. Technical report, National Bureau of Economic Research.

Burstein, A., Cravino, J., and Rojas, M. (2024). Input price dispersion across buyers and misallocation. Technical report, Central Bank of Chile.

De Loecker, J., Eeckhout, J., and Mongey, S. (2021). Quantifying market power and business dynamism in the macroeconomy. Technical report, National Bureau of Economic Research.

Edmond, C., Midrigan, V., and Xu, D. Y. (2023). How costly are markups? *Journal of Political Economy*, 131(7):1619–1675.

Hopenhayn, H. A. (1992). Entry, exit, and firm dynamics in long run equilibrium. *Econometrica: Journal of the Econometric Society*, pages 1127–1150.

Hsieh, C.-T. and Klenow, P. J. (2014). The life cycle of plants in india and mexico. *The Quarterly Journal of Economics*, 129(3):1035–1084.

Melitz, M. J. (2003). The impact of trade on intra-industry reallocations and aggregate industry productivity. *econometrica*, 71(6):1695–1725.

Wilson, R. B. (1993). *Nonlinear pricing*. Oxford University Press, USA.

Appendix

A Optimal Nonlinear Price Derivation

Primitives. Consider a screening problem in which a monopolist offers quantity–transfer bundles to buyers with private productivity types z, drawn from distribution $F(z)$ with density $f(z)$ and support $[\underline{z}, \infty)$. The seller faces constant marginal cost $c > 0$. The buyer's revenue function[22] is $R(z, q)$, increasing in both arguments and differentiable in z. A contract specifies $(q(z), T(z))$, so type z earns net surplus

$$\Pi(z) = R(z, q(z)) - T(z).$$

Seller Problem. The seller chooses a menu $\{q(z), T(z)\}$ to maximize expected profit

$$\max_{\{q(z),T(z)\}} \int_{\underline{z}}^{\infty} \Big[T(z) - cq(z)\Big] f(z)\, dz,$$

subject to individual rationality (IR) and incentive compatibility (IC):

$$\Pi(z) \geq 0, \qquad \Pi(z) \geq R(z, q(\tilde{z})) - T(\tilde{z}) \quad \forall \tilde{z} \geq \underline{z}.$$

We assume monotone allocations $q'(z) \geq 0$, so higher types purchase weakly more.

Envelope and Transfers. By the Envelope Theorem,

$$\Pi'(z) = \frac{\partial R(z, q(z))}{\partial z}, \qquad \Pi(\underline{z}) = 0,$$

so

$$\Pi(z) = \int_{\underline{z}}^{z} \frac{\partial R(s, q(s))}{\partial s}\, ds, \quad T(z) = R(z, q(z)) - \Pi(z).$$

Virtual Surplus. Substituting $T(z)$ into the seller's objective and exchanging the order of integration yields

$$\Pi_{\text{seller}} = \int_{\underline{z}}^{\infty} \Big[R(z, q(z)) - \tfrac{1-F(z)}{f(z)} \tfrac{\partial R(z,q(z))}{\partial z} - cq(z)\Big] f(z)\, dz.$$

Define the virtual surplus

$$\phi(z, q) = R(z, q) - \frac{1}{h(z)} \frac{\partial R(z, q)}{\partial z}, \qquad h(z) = \frac{f(z)}{1 - F(z)},$$

so that

$$\Pi_{\text{seller}} = \int_{\underline{z}}^{\infty} \Big[\phi(z, q(z)) - cq(z)\Big] f(z)\, dz.$$

[22] If buyers are final consumers, $R(z, q)$ can be interpreted as gross utility.

Virtual surplus adjusts revenues for the information rents needed to preserve truthful revelation; the inverse hazard rate $1/h(z)$ scales those rents.

Functional Forms. We now impose the functional forms used in the main text. Let types be Pareto with shape $\kappa > 1$ and, without loss, lower bound $\underline{z} = 1$:

$$f(z) = \kappa z^{-\kappa-1}, \qquad F(z) = 1 - z^{-\kappa}, \qquad h(z) = \frac{\kappa}{z}.$$

Let revenues be homothetic and normalized:

$$R(z, q) = z^{\frac{\sigma-1}{\sigma}} q^{\frac{\sigma-1}{\sigma}}, \qquad \sigma > 1.$$

Then

$$\phi(z, q) = z^{\frac{\sigma-1}{\sigma}} q^{\frac{\sigma-1}{\sigma}} \left(1 - \frac{\sigma - 1}{\kappa\sigma}\right) = z^{\frac{\sigma-1}{\sigma}} q^{\frac{\sigma-1}{\sigma}} \left(\frac{\rho - 1}{\rho}\right), \qquad \rho \equiv \frac{\sigma\kappa}{\sigma - 1}.$$

Allocation. The seller chooses $q(z)$ pointwise:

$$\max_{q(z)} \left\{\frac{\rho - 1}{\rho} z^{\frac{\sigma-1}{\sigma}} q^{\frac{\sigma-1}{\sigma}} - cq\right\}.$$

The FOC gives

$$\frac{\rho - 1}{\rho} \cdot \frac{\sigma - 1}{\sigma} z^{\frac{\sigma-1}{\sigma}} q(z)^{-\frac{1}{\sigma}} = c \quad \Rightarrow \quad q(z) = \left[\frac{1}{c} \frac{\rho - 1}{\rho} \frac{\sigma - 1}{\sigma}\right]^{\sigma} z^{\sigma-1}.$$

This policy is increasing in z and well-behaved whenever $\kappa > \sigma - 1$ (equivalently $\rho > 1$).

Two-Part Tariff Implementation. The allocation can be decentralized by a two-part tariff,

$$T(z) = F + p\, q(z),$$

with per-unit price (marginal price / allocative component)

$$p^{NLP} = \frac{\rho}{\rho - 1} c,$$

and flat fee pinned down by the lowest type's IR:

$$F = R(1, q(1)) - p^{NLP} q(1), \qquad q(1) = \left[\frac{1}{c} \frac{\rho - 1}{\rho} \frac{\sigma - 1}{\sigma}\right]^{\sigma}.$$

Result. Under Pareto-distributed types with parameter $\kappa > \sigma - 1$, homothetic revenue, and constant marginal cost c, the optimal nonlinear price is a two-part tariff

$$T(z) = F + p^{NLP} q(z), \qquad p^{NLP} = \frac{\rho}{\rho - 1} c, \quad \rho = \frac{\sigma\kappa}{\sigma - 1},$$

where F extracts the lowest type's surplus. This mechanism implements the optimal allocation and satisfies IC and IR.

Information-adjusted revenue in closed form. Plugging the optimal $q(z)$ into $\phi(z, q)$ and integrating over types yields

$$\int_1^\infty \phi\big(z, q(z)\big) f(z)\, dz = \frac{\kappa}{\kappa - \sigma + 1} \left(\frac{\sigma\kappa - \sigma + 1}{\sigma\kappa}\right)^{\sigma} \left(\frac{\sigma - 1}{\sigma}\right)^{\sigma - 1} \frac{1}{c^{\sigma - 1}},$$

which exists if and only if $\kappa > \sigma - 1$. This closed form shows the information-adjusted revenue is homogeneous of degree $1 - \sigma$ in marginal cost and depends on primitives only through (σ, κ, c).

Seller revenue is homothetic in q. The virtual surplus can also be rewritten as:

$$\phi(z, q) \ = \ R(z, q) - \frac{1}{h(z)} \frac{\partial R(z, q)}{\partial z} \ = \ \Big(1 - \frac{\sigma - 1}{\sigma\, \kappa}\Big) z^{\frac{\sigma - 1}{\sigma}}\, q^{\frac{\sigma - 1}{\sigma}}.$$

Hence, for any $t > 0$, $\phi(z, tq) = t^{\frac{\sigma - 1}{\sigma}} \phi(z, q)$: the seller-side revenue term is homothetic in q with degree $(\sigma - 1)/\sigma$.

Aggregating over types (support $[1, \infty)$), the seller's revenue functional is

$$\mathcal{R}[q] \ = \ \int_1^\infty \phi\big(z, q(z)\big) f(z)\, dz \ = \ \frac{\sigma\kappa - \sigma + 1}{\sigma} \int_1^\infty z^{\frac{\sigma - 1}{\sigma} - \kappa - 1}\, q(z)^{\frac{\sigma - 1}{\sigma}}\, dz,$$

so for any $t > 0$ we have $\mathcal{R}[tq] = t^{\frac{\sigma - 1}{\sigma}} \mathcal{R}[q]$. This gives the full expression for the seller's revenue and makes its homotheticity in q explicit.

Expected revenue by source Aggregating across buyers, total revenue equals the sum of flat fees and per-unit revenue:

$$\text{Total Revenue} = \underbrace{F \int_1^\infty f(z)\, dz}_{\text{flat fees}} \ + \ \underbrace{\Big(\int_1^\infty q(z)\, f(z)\, dz\Big) p^{\text{NLP}}}_{\text{per-unit revenue}}.$$

With $\int_1^\infty f(z)\,dz = 1$ and $\int_1^\infty z^{\sigma-1} f(z)\,dz = \frac{\kappa}{\kappa-\sigma+1}$ (for $\kappa > \sigma - 1$), and using

$$q(1) = \left[\frac{1}{c}\cdot\frac{\sigma-1}{\sigma}\cdot\frac{\sigma\kappa-\sigma+1}{\sigma\kappa}\right]^{\sigma}, \qquad F = q(1)^{\frac{\sigma-1}{\sigma}} - p^{\text{NLP}}\,q(1), \qquad p^{\text{NLP}} = \frac{\sigma\kappa}{\sigma\kappa-\sigma+1}\,c,$$

Let

$$B \;\equiv\; q(1) = \left[\frac{1}{c}\cdot\frac{\sigma-1}{\sigma}\cdot\frac{\sigma\kappa-\sigma+1}{\sigma\kappa}\right]^{\sigma}, \qquad p^{\text{NLP}} = \frac{\sigma\kappa}{\sigma\kappa-\sigma+1}\,c$$

Total revenue aggregates flat fees and per-unit revenue:

$$\text{Total Revenue} = \underbrace{\left[B^{\frac{\sigma-1}{\sigma}} - p^{\text{NLP}}\,B\right]}_{\text{flat fees}} + \underbrace{p^{\text{NLP}}\,B\,\frac{\kappa}{\kappa-\sigma+1}}_{\text{per-unit revenue}}.$$

The first bracketed term is total flat-fee revenue, while the second term is per-unit markup revenue proportional to the Pareto moment $\mathbb{E}[z^{\sigma-1}] = \kappa/(\kappa-\sigma+1)$; hence heterogeneity ($\kappa$) and technology/costs ($\sigma, c$ through B) shift levels but not the two-part structure.

A.1 Virtual Surplus and Full Participation

The monopolist optimally serves all buyer types, including the lowest type $z = 1$. The logic is transparent in virtual-surplus form. With homothetic revenue,

$$R(z,q) = z^{\frac{\sigma-1}{\sigma}} q^{\frac{\sigma-1}{\sigma}}, \qquad \sigma > 1,$$

and Pareto types with hazard $h(z) = \kappa/z$, we have

$$\frac{\partial R(z,q)}{\partial z} = \frac{\sigma-1}{\sigma} z^{-\frac{1}{\sigma}} q^{\frac{\sigma-1}{\sigma}}, \qquad \phi(z,q) = z^{\frac{\sigma-1}{\sigma}} q^{\frac{\sigma-1}{\sigma}}\Big(1 - \frac{\sigma-1}{\kappa\sigma}\Big) = z^{\frac{\sigma-1}{\sigma}} q^{\frac{\sigma-1}{\sigma}}\,\frac{\rho-1}{\rho}.$$

Hence the net contribution of type z at allocation $q(z)$ is

$$\text{VS}(z) = \phi(z, q(z)) - cq(z).$$

Evaluated at the lowest type,

$$\text{VS}(1) = q(1)^{\frac{\sigma-1}{\sigma}}\,\frac{\rho-1}{\rho} - cq(1),$$

which is strictly positive whenever $\rho > 1$ (equivalently, $\kappa > 1$). Since low types are abundant under Pareto, excluding them lowers profits: the mass of low types more than compensates their low individual surplus. The exclusion trade-off therefore resolves in favor of full participation.

A.2 No Profitable Price Deviation

We show that the seller cannot profit by deviating from the optimal nonlinear schedule and charging a different per-unit price for a given quantity. This validates the two-part tariff with constant unit price $p^{NLP} = \frac{\rho}{\rho-1}c$ as seller-optimal.

A buyer facing marginal price $p(q)$ solves

$$\max_q \{R(z,q) - T(q)\}, \qquad T'(q) = p(q).$$

With $R(z,q) = z^{\frac{\sigma-1}{\sigma}} q^{\frac{\sigma-1}{\sigma}}$,

$$\frac{\partial R}{\partial q} = \frac{\sigma-1}{\sigma} z^{\frac{\sigma-1}{\sigma}} q^{-\frac{1}{\sigma}} = p(q).$$

Solving for the type indifferent at (q,p) yields the inverse demand for the q-th unit:

$$z(q,p) = \left(\frac{\sigma}{\sigma-1}\,p\right)^{\frac{\sigma}{\sigma-1}} q^{\frac{1}{\sigma-1}}.$$

If the seller posts an alternative price p for quantity q, only types $z \geq z(q,p)$ purchase that unit, so demand is

$$D(q,p) = 1 - F\big(z(q,p)\big) = z(q,p)^{-\kappa} \quad \text{(Pareto).}$$

Profit from this deviation is

$$\pi(q,p) = D(q,p)\,(p-c) = \left[\left(\frac{\sigma}{\sigma-1}\,p\right)^{\frac{\sigma}{\sigma-1}} q^{\frac{1}{\sigma-1}}\right]^{-\kappa} (p-c).$$

Maximizing w.r.t. p yields the first-order condition whose solution is

$$\frac{p}{c} = \frac{\rho}{\rho-1}, \qquad \rho = \frac{\sigma\kappa}{\sigma-1},$$

i.e., $p = p^{NLP}$. Any unilateral price deviation reduces profit. Thus the nonlinear price is robust to such deviations and coincides with the allocative price embedded in the mechanism (cf. Wilson (1993)).

Figure A1: No Profitable Price Deviation

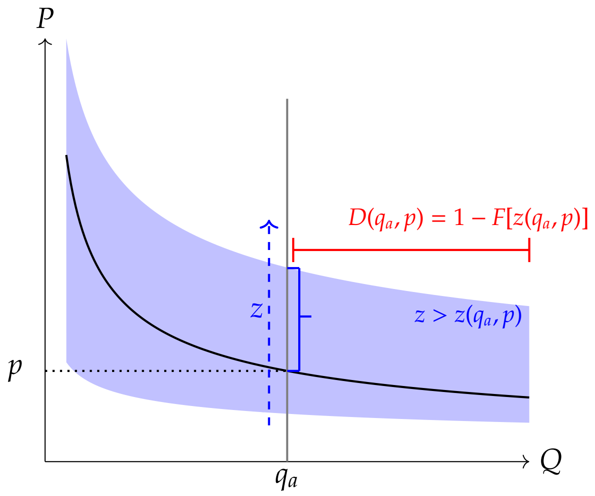


Figure A1 illustrates the logic: raising p at a given q_a shrinks the set of buyers above the new cutoff $z(q_a, p)$; the loss in volume offsets the price gain unless $p = p^{NLP}$.

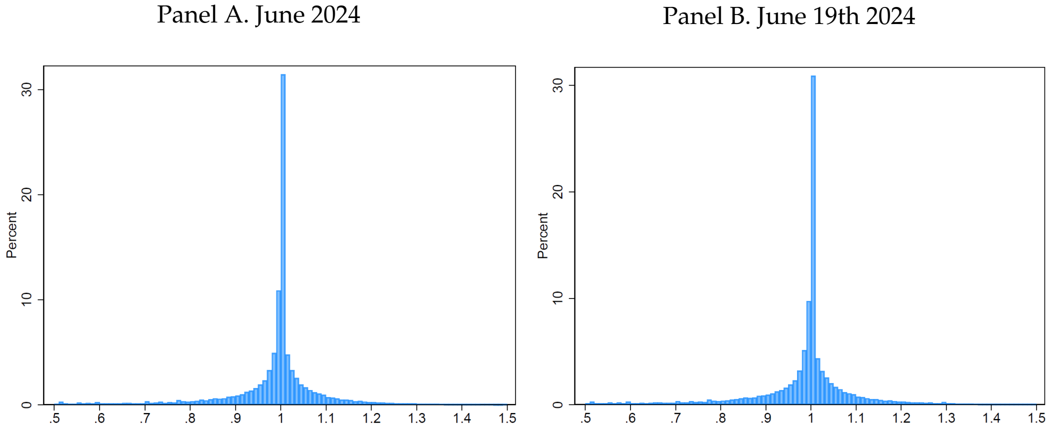
B Additional Descriptive Evidence

B.1 Price Dispersion Histograms

We observe substantial price variations for a given seller i and product g (the "detail" variable in the invoice) within a month. Following Burstein et al. (2024), we construct a price dispersion measure $\tilde{p}_{igt}$ for June 2024, the month with the most transactions in 2024. We divide unit prices observed for each product g transaction from seller i to buyer j by the mean price across seller i and product g. We repeat the same exercise for June 19th, 2024, the day with the most transactions in that month, to ensure our results are not driven by month-specific demand and supply shocks. The variance of $\ln(\alpha_{ijg})$ is 0.65 monthly and 0.61 daily, and 29% of transactions in both cases show no price dispersion.

in Figure A2. The histogram does not vary substantially from monthly to daily basis, while supply and demand shocks could still explain price differences. For 71% of transactions, we cannot reject that firms engage in some form of price discrimination departing from uniform pricing, although none of the exercises in this section aim to be causal, but rather describe equilibrium objects observed in the data and test which variables they correlate with in search of indicative evidence.

Figure A2: Price dispersion



Notes: The figure reports the distribution of the log of demeaned price for the month of June 2024. We exclude seller-product pairs with only one transaction.

B.2 Residual Price Determinants by Selected Industries

To assess whether the pattern of nonlinear pricing driven by buyer observables generalizes across sectors, we replicate the residual decomposition analysis for the two industries with the highest volume of transactions: Manufacturing and Retail and Wholesale. For both sectors, we estimate the following regression:

$$\ln p_{igjt} = \beta_0 + \Psi_{igmS} + \epsilon_{ijgt}, \tag{21}$$

where p_{igjt} is the unit price of a product g sold by seller i to buyer j at time t, and Ψ_{igmS} represents seller-product-month fixed effects interacted with different sets of quantity and buyer-side controls S. Buyer groups B are defined based on 11 sectors, 3 firm-size categories, and 16 regions.

Table A1: Price residual determinants: Manufacturing

	(1)	(2)	(3)	(4)
R^2	0.581	0.389	0.312	0.776
S = Quantity	✓			
S = Buyer		✓		
S = Buyer Group			✓	
S = Quantity × Buyer Group				✓
N	136M	136M	136M	136M

Table A2: Price residual determinants: Retail and Wholesale

	(1)	(2)	(3)	(4)
R^2	0.296	0.391	0.309	0.471
S = Quantity	✓			
S = Buyer		✓		
S = Buyer Group			✓	
S = Quantity × Buyer Group				✓
N	180M	180M	180M	180M

In both sectors, the pattern remains unchanged: quantity discounts (second-degree price discrimination) and buyer group-based pricing (third-degree) explain the majority of price dispersion once product and time effects are controlled for. This reinforces our main finding that nonlinear prices shaped by buyer-side observables are a pervasive feature of pricing in supply chains.

B.3 Average Quantity Discount by Sector

Table A3: Average Quantity Discount by Sector

Sector	Mean Q discount	N transactions
All sectors	-0.042	430M
Agriculture	-0.042	2M
Mining	-0.016	1M
Manufacturing	-0.036	118M
Utilities	0.000	6M
Construction	-0.129	1M
Retail and Wholesale	-0.048	270M
Transport & ICTs	-0.032	12M
Financial Services	-0.002	49M
Real Estate Services	-0.052	1M
Business Services	-0.089	5M
Personal Services	-0.053	1M

B.4 Test for Buyer Power Data Generation Process

To examine whether observed quantity discounts reflect buyer power rather than seller-driven price discrimination, we exploit cross-sectional variation in the number of suppliers each buyer transacts with during the sample period. The underlying idea is that buyers with access to a larger number of sellers may possess stronger outside options, enhancing their bargaining position and enabling them to negotiate better pricing terms. We define buyer power as the logarithm of the total number of distinct sellers each buyer purchases from within the observed month. We then test whether buyer power flattens quantity discounts by estimating the interaction between log quantity and buyer power in a log-linear price regression. Specifically, we estimate:[23]

$$\ln p_{igjt} = \beta_0 + \beta_1 \ln q_{igjt} + \beta_2 \left(\log q_{ijt} \times \log \text{NumProviders}_j\right) + \Psi_{igm} + \epsilon_{ijgt},$$

A positive coefficient on the interaction term ($\beta_2 > 0$) would suggest that quantity discounts become flatter as buyer power increases, consistent with buyers using their broader supplier base to resist steep discounts or nonlinear price schedules.

[23] Standard errors are clustered at the buyer level to account for within-buyer correlation.

We find that $\beta_1 = -0.0462$ and $\beta_2 = -0.0098$, both estimated with standard errors below 0.0001. While the interaction term is statistically significant, the magnitude is economically negligible. This suggests that buyer power, as measured by the number of suppliers, does not appear to be the primary mechanism generating quantity discounts. If anything, the evidence is more consistent with seller-driven price discrimination rather than buyer power shaping quantity discounts.

B.5 Firm Sales Partition

We find that firms in Chile have a clear partition on firms' buyers: 79% of firms weighted by sales sell all their output either to only other firms (67%) or to only final consumers (12%). As we can combine firm-to-firm transaction data with firms' accounting variables, we build an indicator variable that takes the value of 0 if all firm sales go to final consumers and 1 if sales go only to other firms, and we weigh the indicator by firm sales.

Table A4: Firms sales partition

Sector (Supply Chain Transactions Value Share)	**All to final consumer**	**All to other firms**
Firm population (100%)	0.12	0.67
Agriculture (2%)	0.05	0.60
Mining (1%)	0.27	0.08
Manufacturing (15%)	0.06	0.69
Utilities (3%)	0.20	0.52
Construction (8%)	0.02	0.89
Retail and Wholesale (32%)	0.09	0.69
Transport and ICTs (10%)	0.16	0.68
Financial Services (18%)	0.18	0.68
Real Estate Services (1%)	0.25	0.38
Business Services (7%)	0.09	0.81
Personal Services (2%)	0.69	0.10

Notes: Exports are excluded. The remaining 16% of sales shares for the firm population are firms that sell to both final consumers and other firms. We observe firm-to-firm sales an fims total sales, we compute the sales to consumer as the residual between both. For 2% of firms, we get negative sales to consumers and exclude them from this table.

As shown in Table A4, there is heterogeneity across sectors, though the partition between firms selling to final consumers and other firms is present across all sectors.

C Model Details and Derivations

C.1 Verification for Retailers

Within retail sector s, CES demand over differentiated varieties implies

$$y_j \;=\; Y_s\Big(\frac{p_j}{P_s}\Big)^{-\varphi_s}, \qquad Y_s = \theta_s Y, \quad P_Y \equiv 1,$$

hence the inverse demand $p_j = P_s\,(y_j/Y_s)^{-1/\varphi_s}$. Revenue as a function of own output Q_j is

$$R_j \;=\; p_j Q_j \;=\; P_s\,(\theta_s Y)^{1/\varphi_s}\,Q_j^{(\varphi_s-1)/\varphi_s}.$$

Therefore, for retailers ($\ell = r$) the revenue guess for retailers holds with

$$\psi_s^r = \frac{\varphi_s - 1}{\varphi_s}, \qquad A_s^r = P_s\,(\theta_s Y)^{1/\varphi_s}.$$

C.2 Verification for Upstream Sellers (Homothetic Revenue)

Fix a seller sector s' with elasticity $\sigma_{s'} > 1$ and a seller $j \in \mathcal{U}_{s'}$ with marginal cost $c_j > 0$. For buyers in partition (ℓ, s), the two–part tariff of Proposition 1 implies the allocative price $p_{js}^{\ell} = \mu_{ss'}^{\ell} c_j$ with $\mu_{ss'}^{\ell} = \rho_{ss'}^{\ell}/(\rho_{ss'}^{\ell} - 1)$ and $\rho_{ss'}^{\ell} = \xi_s^{\ell}\sigma_{s'} > \sigma_{s'}$.

Step 1: quantity aggregation. By the CES share rule, for buyer $i = (\ell, s, z_i)$,

$$m_{ij} \;=\; M_{is'}\left(\frac{p_{js}^{\ell}}{P_{ss'}^{\ell}}\right)^{-\sigma_{s'}}.$$

Let $\tilde{\nu}_{\ell s}$ be the buyer distribution in (ℓ, s) normalized to one and define the average sector–s' bundle per buyer

$$\widehat{D}_{ss'}^{\ell} \;\equiv\; \int M_{is'}\, d\tilde{\nu}_{\ell s}(i).$$

Total quantity sold by j to partition (ℓ, s) is then

$$Q_j^{\ell,s} \;=\; N_s^{\ell}\,\widehat{D}_{ss'}^{\ell}\left(\frac{p_{js}^{\ell}}{P_{ss'}^{\ell}}\right)^{-\sigma_{s'}}.$$

Summing over (ℓ, s) and substituting $p_{js}^{\ell} = \mu_{ss'}^{\ell} c_j$,

$$Q_j \;=\; \sum_{\ell,s} N_s^{\ell}\, \widehat{D}_{ss'}^{\ell} \left(\frac{\mu_{ss'}^{\ell} c_j}{P_{ss'}^{\ell}} \right)^{-\sigma_{s'}} \;=\; c_j^{-\sigma_{s'}} \underbrace{\sum_{\ell,s} N_s^{\ell}\, (\mu_{ss'}^{\ell})^{-\sigma_{s'}} (P_{ss'}^{\ell})^{\sigma_{s'}}\, \widehat{D}_{ss'}^{\ell}}_{\mathcal{D}_{s'}} . \tag{22}$$

Hence

$$c_j^{1-\sigma_{s'}} \;=\; \left(\tfrac{Q_j}{\mathcal{D}_{s'}} \right)^{\frac{\sigma_{s'}-1}{\sigma_{s'}}} . \tag{23}$$

Step 2: variable (marginal–price) revenue. Revenue at the allocative margin from partition (ℓ, s) is

$$R_j^{\mathrm{lin};\,\ell,s} \;=\; \int p_{js}^{\ell} m_{ij}\, d\nu_{\ell s}(i) \;=\; N_s^{\ell}\, \widehat{D}_{ss'}^{\ell}\, (\mu_{ss'}^{\ell} c_j)^{1-\sigma_{s'}} (P_{ss'}^{\ell})^{\sigma_{s'}-1}.$$

Summing across partitions and using (23),

$$R_j^{\mathrm{lin}} \;=\; \left(\tfrac{Q_j}{\mathcal{D}_{s'}} \right)^{\frac{\sigma_{s'}-1}{\sigma_{s'}}} \underbrace{\sum_{\ell,s} N_s^{\ell}\, (\mu_{ss'}^{\ell})^{1-\sigma_{s'}} (P_{ss'}^{\ell})^{\sigma_{s'}-1}\, \widehat{D}_{ss'}^{\ell}}_{\mathcal{S}_{s'}} . \tag{24}$$

Step 3: flat–fee revenue. For partition (ℓ, s), the fee for the lowest buyer type $\underline{z}_s^{\ell}$ satisfies (cf. Proposition)

$$F_{js}^{\ell}(\underline{z}_s^{\ell}) \;=\; \frac{1}{\sigma_{s'} - 1}\, P_{ss'}^{\ell}\, M_{is'}(\underline{z}_s^{\ell}) \left(\frac{p_{js}^{\ell}}{P_{ss'}^{\ell}} \right)^{1-\sigma_{s'}} .$$

Aggregating over buyers in (ℓ, s) yields

$$R_j^{\mathrm{fee};\,\ell,s} \;=\; N_s^{\ell}\, \frac{P_{ss'}^{\ell} M_{is'}(\underline{z}_s^{\ell})}{\sigma_{s'} - 1}\, (\mu_{ss'}^{\ell} c_j)^{1-\sigma_{s'}} (P_{ss'}^{\ell})^{\sigma_{s'}-1}.$$

Summing across partitions and using (23),

$$R_j^{\mathrm{fee}} \;=\; \left(\tfrac{Q_j}{\mathcal{D}_{s'}} \right)^{\frac{\sigma_{s'}-1}{\sigma_{s'}}} \underbrace{\sum_{\ell,s} N_s^{\ell}\, (\mu_{ss'}^{\ell})^{1-\sigma_{s'}} (P_{ss'}^{\ell})^{\sigma_{s'}-1}\, \frac{P_{ss'}^{\ell} M_{is'}(\underline{z}_s^{\ell})}{\sigma_{s'} - 1}}_{\mathcal{F}_{s'}} . \tag{25}$$

Step 4: revenue representation (homotheticity) and closed–form scale. Adding (24) and (25),

$$R_j \;=\; \left(\tfrac{Q_j}{\mathcal{D}_{s'}} \right)^{\frac{\sigma_{s'}-1}{\sigma_{s'}}} (\mathcal{S}_{s'} + \mathcal{F}_{s'}) \;=\; A_{s'}^{u}\, Q_j^{(\sigma_{s'}-1)/\sigma_{s'}}, \tag{26}$$

with

$$A^u_{s'} \;=\; \mathcal{D}_{s'}^{-\frac{\sigma_{s'}-1}{\sigma_{s'}}}\,(\mathcal{S}_{s'}+\mathcal{F}_{s'}), \qquad \begin{aligned} \mathcal{D}_{s'} &\equiv \sum_{\ell,s} N^\ell_s\,(\mu^\ell_{ss'})^{-\sigma_{s'}}(P^\ell_{ss'})^{\sigma_{s'}}\,\widehat{D}^\ell_{ss'}, \\ \mathcal{S}_{s'} &\equiv \sum_{\ell,s} N^\ell_s\,(\mu^\ell_{ss'})^{1-\sigma_{s'}}(P^\ell_{ss'})^{\sigma_{s'}-1}\,\widehat{D}^\ell_{ss'}, \\ \mathcal{F}_{s'} &\equiv \sum_{\ell,s} N^\ell_s\,(\mu^\ell_{ss'})^{1-\sigma_{s'}}(P^\ell_{ss'})^{\sigma_{s'}-1}\frac{P^\ell_{ss'}M_{is'}(\underline{z}^\ell_s)}{\sigma_{s'}-1}. \end{aligned} \tag{27}$$

Thus, upstream revenue is homogeneous of degree $\psi^u_{s'} = (\sigma_{s'}-1)/\sigma_{s'}$ in own output, with shifter $A^u_{s'}$ comprising a variable (marginal–price) component $\mathcal{S}_{s'}$ and a flat–fee component $\mathcal{F}_{s'}$, both attenuated by the effective demand index $\mathcal{D}_{s'}$.

C.3 Kuhn–Tucker implementation of marginal–cost pricing

Following the logic of Theorem 1 in Baqaee and Farhi (2020), consider a social planner who chooses final consumption C, final demands $\{y_j\}$, outputs $\{Q_j\}$, intermediate allocations $\{m_{ij}\}$, and masses of active producers $\{N_j\}$ to maximize $U(C)$ subject to (i) material balance for each variety j, (ii) per–producer feasibility with mass N_i of active buyers/producers i, and (iii) entry costs $\mathcal{K}_i$ paid in units of the final good:

$$Q_j - y_j - \sum_i N_i\,m_{ij} = 0, \qquad Q_i \le N_i\,\mathcal{F}_i(m_{i\cdot}), \qquad C \ge \sum_i \mathcal{K}_i N_i, \qquad \sum_j y_j = C.$$

Form the Lagrangian with Kuhn–Tucker multipliers v_j (material balance for good j), η_i (feasibility for producer i), and ϑ (final–good/entry resource):

$$\mathcal{L} = U(C) + \sum_j v_j\left(Q_j - y_j - \sum_i N_i m_{ij}\right) + \sum_i \eta_i(N_i\mathcal{F}_i(m_{i\cdot}) - Q_i) + \vartheta\left(C - \sum_i \mathcal{K}_i N_i\right).$$

FOCs and complementary slackness:

$$\frac{\partial\mathcal{L}}{\partial m_{ij}}: \quad -v_j N_i + \eta_i N_i\,\frac{\partial\mathcal{F}_i}{\partial m_{ij}} = 0 \;\Rightarrow\; v_j = \eta_i\,\frac{\partial\mathcal{F}_i}{\partial m_{ij}} \quad (\forall i, j).$$

$$\frac{\partial\mathcal{L}}{\partial Q_j}: \quad v_j - \eta_j = 0 \;\Rightarrow\; v_j = \eta_j \quad (\forall j).$$

$$\frac{\partial\mathcal{L}}{\partial y_j}: \quad -v_j + \lambda = 0 \;\Rightarrow\; v_j = \lambda \quad (\forall j), \quad \text{with } \lambda \text{ the multiplier on } \sum_j y_j = C.$$

$$\frac{\partial\mathcal{L}}{\partial C}: \quad U'(C) - \lambda + \vartheta = 0.$$

$$\frac{\partial \mathcal{L}}{\partial N_i}: \quad -\sum_j v_j m_{ij} + \eta_i \mathcal{F}_i(m_{i\cdot}) - \vartheta\, \mathcal{K}_i \le 0, \quad N_i \ge 0, \quad N_i\Big(-\sum_j v_j m_{ij} + \eta_i \mathcal{F}_i - \vartheta \mathcal{K}_i\Big) = 0.$$

Using $v_j = \eta_j$, the input FOCs become

$$\eta_j \;=\; \eta_i \frac{\partial \mathcal{F}_i}{\partial m_{ij}} \quad (\forall i, j),$$

which are the planner's cost–minimizing conditions: effective prices (proportional to η_j) equal marginal costs along every input link. The entry condition states that, at the optimum, the surplus created by an additional firm i (valued at $\eta_i \mathcal{F}_i - \sum_j \eta_j m_{ij}$) equals the entry cost valued at $\vartheta \mathcal{K}_i$ whenever $N_i > 0$.

Decentralized implementation with markups, rebates, and entry. Suppose in the decentralized economy each seller j charges a constant markup $\mu_j > 1$ on intermediate sales, and retailers in buyer sector s charge $\mu_s^r > 1$ to final consumers. Introduce ad valorem rebates on purchases so buyers face effective marginal prices

$$\tilde{p}_{ij} = (1 - t_j)\, p_{ij}, \qquad t_j \equiv 1 - \frac{1}{\mu_j}, \qquad \tilde{p}_s = (1 - t_s^r)\, p_s, \qquad t_s^r \equiv 1 - \frac{1}{\mu_s^r}.$$

Then $\tilde{p}_{ij} = c_j$ for intermediates and $\tilde{p}_s = \mathrm{MC}_s$ for retail. Firm i's cost minimization with $\tilde{p}_{ij}$ yields

$$\tilde{p}_{ij} = \widetilde{\eta}_i \frac{\partial \mathcal{F}_i}{\partial m_{ij}} \;\Rightarrow\; c_j = \widetilde{\eta}_i \frac{\partial \mathcal{F}_i}{\partial m_{ij}},$$

which coincides with the planner's FOCs after a common normalization of shadow values ($\widetilde{\eta}_i \propto \eta_i$). Hence the decentralized $\{m_{ij}\}$, $\{Q_j\}$, and C replicate the planner's allocation. Two–part tariffs' flat fees are infra–marginal and do not affect these FOCs.

Financing and entry. Let each active firm i pay a non–distortionary license $\mathcal{T}_i$ and let the government rebate $t_j p_{ij} m_{ij}$ and $t_s^r p_s y_s$ to buyers. Setting $\mathcal{T}_i = \vartheta\, \mathcal{K}_i$ ensures that the decentralized free–entry condition (operating profits net of input rebates minus the license equal zero) matches the planner's complementary slackness for N_i. Because licenses and flat fees are infra–marginal, they do not alter marginal conditions, while markups can remain strictly positive to fund entry costs. Government budget balance is achieved by choosing $\{\mathcal{T}_i\}_i$ to equal the present value of rebate outlays at the implemented allocation.

Therefore, the planner's allocation can be implemented even when firms charge markups: purchase–side rebates neutralize marginal wedges so that buyers face marginal costs in all nests, and entry costs are financed by non–distortionary fixed charges, preserving the planner's first–order conditions and entry margins.

C.4 Seller–identity invariance of the total unit markup

By the CES share rule within seller sector s', buyer i's demand for seller j's variety is

$$m_{ij} \;=\; M_{is'} \left(\frac{p^{\ell}_{js}}{P^{\ell}_{ss'}} \right)^{-\sigma_{s'}} .$$

From the optimal tariff characterization, the flat fee charged to buyer i in partition (ℓ, s) by seller $j \in \mathcal{U}_{s'}$ (with $\sigma_{s'} > 1$) is

$$F^{\ell}_{js} \;=\; \frac{1}{\sigma_{s'} - 1}\, \tau_{is'}(\underline{z}^{\ell}_{s}) \left(M_{is'}(\underline{z}^{\ell}_{s}) \right)^{\frac{\sigma_{s'}-1}{\sigma_{s'}}} \left(\frac{p^{\ell}_{js}}{P^{\ell}_{ss'}} \right)^{1-\sigma_{s'}}, \qquad \tau_{is'} \equiv P^{\ell}_{ss'} M^{1/\sigma_{s'}}_{is'} .$$

Divide the fee by quantity using the share rule:

$$\frac{F^{\ell}_{js}}{m_{ij}} = \frac{p^{\ell}_{js}}{P^{\ell}_{ss'}} \cdot \frac{\frac{1}{\sigma_{s'}-1}\, \tau_{is'}(\underline{z}^{\ell}_{s}) \left(M_{is'}(\underline{z}^{\ell}_{s}) \right)^{\frac{\sigma_{s'}-1}{\sigma_{s'}}}}{M_{is'}} .$$

Total unit price is $\frac{T_{ij}}{m_{ij}} = p^{\ell}_{js} + \frac{F^{\ell}_{js}}{m_{ij}}$. Using $p^{\ell}_{js} = \mu^{\ell}_{ss'} c_j$ and dividing by c_j yields

$$\frac{\frac{T_{ij}}{m_{ij}}}{c_j} = \mu^{\ell}_{ss'} \left[1 + \frac{1}{P^{\ell}_{ss'}} \cdot \frac{\frac{1}{\sigma_{s'}-1}\, \tau_{is'}(\underline{z}^{\ell}_{s}) \left(M_{is'}(\underline{z}^{\ell}_{s}) \right)^{\frac{\sigma_{s'}-1}{\sigma_{s'}}}}{M_{is'}} \right] \;\equiv\; \mu^{\ell}_{ss'} \big(1 + \chi^{\ell}_{ss'}(i) \big).$$

The term $\chi^{\ell}_{ss'}(i)$ depends only on buyer–side objects within (ℓ, s) and on the sectoral index $P^{\ell}_{ss'}$, but not on seller j. Therefore the total unit markup is invariant to the seller's identity within a given buyer partition.

C.5 Profit Functions Results

Profits of the Lowest Type. For all retail sectors s, equilibrium profits of the lowest–productivity firm $i = (r, s, \underline{z}^{r}_{s})$ are not necessarily zero, nor necessarily positive. In particular,

$$\Pi(\underline{z}^{r}_{s}) > 0 \quad \iff \quad \frac{1}{(1 - \alpha^{r}_{s})(\varphi_s - 1)} \;\geq\; \zeta_s, \qquad \zeta_s := \sum_{s' \in \mathcal{S}} \frac{\theta^{r}_{ss'}}{\sigma_{s'} - 1}.$$

Consider a retail firm $i = (r, s, z_i)$ with $z_i = \underline{z}^{r}_{s}$. Denote by *variable profits* the component net of flat fees. Since retailers charge a constant markup $\varphi_s/(\varphi_s - 1)$, variable profits equal a fixed share of revenue:

$$\mathrm{VarProf}(\underline{z}_s^r) \;=\; \frac{1}{\varphi_s}\,\mathrm{Revenue}(\underline{z}_s^r).$$

Total profits subtract the flat fees paid to upstream suppliers,

$$\Pi(\underline{z}_s^r) \;=\; \mathrm{VarProf}(\underline{z}_s^r) \;-\; \sum_{s'\in\mathcal{S}} F^r_{ss'}\,N^u_{s'}.$$

Substituting yields

$$\Pi(\underline{z}_s^r) \;=\; \frac{\mathrm{Revenue}(\underline{z}_s^r)}{\varphi_s}\Big[1-(1-\alpha_s^r)(\varphi_s-1)\,\zeta_s\Big].$$

Note that ζ_s is a weighted average of $\frac{1}{\sigma_{s'}-1}$ across seller sectors, with weights given by the Cobb–Douglas elasticities $\theta^r_{ss'}$, which satisfy $\sum_{s'}\theta^r_{ss'}=1$. Thus, more important inputs in production receive greater weight in ζ_s.

Average profits. It is useful to restate how revenue scales with productivity. For retail and upstream firms we have

$$\mathrm{Revenue}(z_s^r) = \left(\frac{z_s^r}{\widetilde{z}_s^r}\right)^{\varphi_s-1}\mathrm{Revenue}(\underline{z}_s^r), \qquad \mathrm{Revenue}(z_{s'}^u) = \left(\frac{z_{s'}^u}{\widetilde{z}_{s'}^u}\right)^{\sigma_{s'}-1}\mathrm{Revenue}(\underline{z}_{s'}^u).$$

This scaling implies that average profits can be expressed in closed form.

For retailers,

$$\begin{aligned}
\mathbb{E}\,[\Pi_s^r] &= \mathrm{Revenue}(\widetilde{z}_s^r) - \sum_{s'\in\mathcal{S}} F^r_{ss'}\,N^u_{s'} \\
&= \frac{\mathrm{Revenue}(\underline{z}_s^r)}{\varphi_s}\left[\left(\frac{\widetilde{z}_s^r}{\underline{z}_s^r}\right)^{\varphi_s-1} - (1-\alpha_s^r)(\varphi_s-1)\sum_{s'\in\mathcal{S}}\frac{\theta^r_{ss'}}{\sigma_{s'}-1}\right].
\end{aligned}$$

For $j\in\mathcal{U}_{s'}$, expected profits decompose into (i) variable profits from the allocative margin, (ii) flat fees collected from buyers, and (iii) flat fees paid upstream:

$$\mathbb{E}\left[\Pi_j^u\right] = \underbrace{\sum_{\ell\in\{u,r\}}\sum_{s\in\mathcal{S}}\int_{i\in\mathcal{F}_{\ell s}}\left(p^\ell_{js}-c_j\right)m_{ij}\,d\nu_{\ell s}(i)}_{\text{variable profits from allocative margin}} \;+\; \underbrace{\sum_{\ell\in\{u,r\}}\sum_{s\in\mathcal{S}}\int_{i\in\mathcal{F}_{\ell s}}F^\ell_{js}(i)\,d\nu_{\ell s}(i)}_{\text{flat–fee revenue collected}} \;-\; \underbrace{\sum_{t\in\mathcal{S}}\int_{h\in\mathcal{U}_t}F^u_{hs'}(j)\,d\nu_t(h)}_{\text{flat–fee payments}}.$$

Here p^ℓ_{js} is the allocative (marginal) price charged by seller j to buyers in partition (ℓ,s), c_j is seller j's marginal cost, m_{ij} is buyer i's quantity purchased from j, $F^\ell_{js}(i)$ is the flat fee i pays to j, $\mathcal{F}_{\ell s}$ is the set of active buyers of type ℓ in sector s with measure $\nu_{\ell s}$, and the last term aggregates the flat fees $F^u_{hs'}(j)$ that j pays to its own upstream suppliers $h\in\mathcal{U}_t$.

Average upstream profits depend only on sectoral labor allocation. Fix technology and demand primitives $\{\alpha, \theta, \sigma\}$ and the productivity distributions (so that $\{\underline{z}, \widetilde{z}\}$ are fixed). Then the expected profit of the average upstream firm in sector s' depends only on sectoral labor allocation according to:

$$\mathbb{E}\left[\Pi^u_{s'}\right] = \frac{1}{N^u_{s'}}\left[\sum_{s \in \mathcal{S}} w\, L^r_s\, \Lambda^r_{ss'} \;+\; \sum_{t \in \mathcal{S}} w\, L^u_t\, \Lambda^u_{ts'}\right] \;-\; w\, l^u_{s'} \sum_{t \in \mathcal{S}} \Lambda^u_{s't},$$

where $L^\ell_s = l^\ell_s(\widetilde{z}^\ell_s)\, N^\ell_s$ is total labor used in sector (ℓ, s), and $l^\ell_s(\widetilde{z}^\ell_s)$ denotes labor of the average variety (productivity $\widetilde{z}^\ell_s$). The coefficients are

$$\Lambda^r_{ss'} \;=\; \frac{(1-\alpha^r_s)\, \theta^r_{ss'}}{\alpha^r_s}\left(\frac{1}{\sigma_{s'}-1}\Big(\tfrac{\underline{z}^r_s}{\widetilde{z}^r_s}\Big)^{\varphi_s - 1} + 1\right), \qquad \Lambda^u_{s't} \;=\; \frac{(1-\alpha^u_{s'})\, \theta^u_{s't}}{\alpha^u_{s'}}\left(1 + \frac{1}{\sigma_t - 1}\Big(\tfrac{\underline{z}^u_{s'}}{\widetilde{z}^u_{s'}}\Big)^{\sigma_{s'}-1}\right).$$

Hence, conditional on primitives, $\mathbb{E}\Pi^u_{s'}$ varies solely with the sectoral labor aggregates $\{w\, L^\ell_s\}$ and own $w\, l^u_{s'}$[24].

C.6 Welfare decomposition

This appendix provides a complete and notation-consistent derivation of Proposition 2. We proceed in six steps: (i) normalization and free entry; (ii) notation and dimensions; (iii) sectoral indices and the upstream marginal-cost recursion; (iv) stacking and the linear-systems representation; (v) exposure maps and aggregation of buyer-specific wedges with the equivalence lemma; (vi) selection invariance.

Normalization and free entry. We work in steady state and normalize the household's labor endowment to one. Let $\delta \in (0,1)$ be the per-period exit probability and m the mass of entrants. The law of motion implies $N = m/(1-\delta)$ at the sector level and in aggregate. Free entry equates expected discounted profits to entry costs in wage units, so with entry cost c_e,

$$\frac{\mathbb{E}[\pi]}{1-\delta} = w\, c_e \quad\Rightarrow\quad \mathbb{E}[\pi] = (1-\delta)\, w\, c_e.$$

Aggregate operating profits are $\Pi = N\, \mathbb{E}[\pi] = \frac{m}{1-\delta} \cdot (1-\delta)\, w\, c_e = m\, w\, c_e$. Nominal income is $Y = w\, L_{\text{prod}} + \Pi = w\big(L_{\text{prod}} + m\, c_e\big)$. With unit labor endowment, $L_{\text{prod}} + m\, c_e = 1$, hence $Y = w$. Because indirect utility is homogeneous of degree zero in prices and income, measuring utility in wage units yields $W = \frac{Y}{P_Y} = \frac{w}{P_Y} = \frac{1}{P_Y}$.

[24] Sketch. Start from the upstream profits decomposition . Use (i) p–P CES shares and the identity $(Q^u_{s'}/D_{s'})^{(\sigma_{s'}-1)/\sigma_{s'}} = 1/N^u_{s'}$, (ii) the flat–fee formula $F^\ell_{ss'} = \frac{P^\ell_{ss'} M^\ell_{ss'}(\underline{z}^\ell_s)}{N^u_{s'}(\sigma_{s'}-1)}$, and (iii) cost minimization $P^\ell_{ss'} M^\ell_{ss'} = w\, l^\ell_s \frac{(1-\alpha^\ell_s)\theta^\ell_{ss'}}{\alpha^\ell_s}$ to aggregate across buyers and suppliers. Collecting terms yields the stated affine function of $\{w\, L^\ell_s\}$ and $w\, l^u_{s'}$ with coefficients Λ.

Notation and dimensions. We use lower-case symbols for logarithms, taken componentwise for vectors and elementwise for matrices. For any square matrix $X = [x_{ij}]_{i,j=1}^{S} \in \mathbb{R}^{S\times S}$, define the column-stacking operator

$$\operatorname{vec}(X) = \begin{bmatrix} x_{11} \\ \vdots \\ x_{S1} \\ x_{12} \\ \vdots \\ x_{S2} \\ \vdots \\ x_{1S} \\ \vdots \\ x_{SS} \end{bmatrix} \in \mathbb{R}^{S^2}.$$

Let $\mathbf{1}_m$ be the $m \times 1$ vector of ones, e_i the i-th canonical basis vector, $\operatorname{diag}(v)$ the diagonal matrix formed from a vector v, and $\otimes$ the Kronecker product.

Objects and dimensions.

Object	Dimension and meaning
$\theta \in \mathbb{R}^{\lvert\mathcal{S}\rvert}$	final-expenditure shares (used as row $\theta^\top$)
A^{uu}, $B^{ru} \in \mathbb{R}^{\lvert\mathcal{S}\rvert\times\lvert\mathcal{S}\rvert}$	cost-share matrices: upstream–upstream and retail–upstream
$\tilde{\lambda}_{ru}$, $\tilde{\lambda}_u \in \mathbb{R}^{1\times\lvert\mathcal{S}\rvert}$	final-demand exposures (row vectors), cf. (20)
μ^r, $\mu^{uu} \in \mathbb{R}^{\lvert\mathcal{S}\rvert}$	sector-level wedges: upstream to retail, upstream to upstream
N^r, $N^u \in \mathbb{R}^{\lvert\mathcal{S}\rvert}$	masses of active varieties: retail and upstream
$\mathcal{V}$, $\mathcal{V}^u \in \mathbb{R}^{\lvert\mathcal{S}\rvert}$	selection terms: retail and upstream
φ, $\sigma \in \mathbb{R}^{\lvert\mathcal{S}\rvert}$	CES elasticities by sector (retail and upstream)
$C^u \in \mathbb{R}^{\lvert\mathcal{S}\rvert}$	upstream marginal cost indices

Division by $(\sigma - 1)$ is elementwise, and all $\Delta\log(\cdot)$ act componentwise.

Sectoral indices and the upstream marginal-cost recursion. The retail sector-s price index is

$$P_s = \mu_s^r\,\Theta_s^r\,w^{\alpha_s^r}\left(\prod_{s'\in\mathcal{S}}\left(P_{s's}^r\right)^{(1-\alpha_s^r)\,\theta_{ss'}^r}\right)\left(N_s^r\right)^{-\frac{1}{\varphi_s-1}}\mathcal{V}_s, \qquad P_{s's}^r = \mu_{s's}^r\,C_{s'}. \tag{28}$$

For each upstream sector s', marginal cost obeys

$$C_{s'} = \Theta^u_{s'}\, w^{\alpha^u_{s'}} \left(\prod_{v\in\mathcal{S}} \left(P^u_{vs'}\right)^{(1-\alpha^u_{s'})\,\theta^u_{s'v}}\right) \left(N^u_{s'}\right)^{-\frac{1}{\sigma_{s'}-1}} \mathcal{V}^u_{s'}, \qquad P^u_{vs'} = \mu^u_{vs'}\, C_v. \tag{29}$$

Taking logs yields

$$\log C_{s'} = \sum_v (1-\alpha^u_{s'})\,\theta^u_{s'v}\left(\log\mu^u_{vs'} + \log C_v\right) + \alpha^u_{s'}\log w + \log\Theta^u_{s'} - \frac{1}{\sigma_{s'}-1}\log N^u_{s'} + \log\mathcal{V}^u_{s'}. \tag{30}$$

Stack (30) across s' and set $A^{uu}_{s'v} := (1-\alpha^u_{s'})\,\theta^u_{s'v}$. Then

$$\log C^u = A^{uu}\,\log C^u + \log\mu^{uu} + \alpha^u\log w + \log\Theta^u - \frac{\log N^u}{\sigma-1} + \log\mathcal{V}^u. \tag{31}$$

For differences between two equilibria, technology constants and the wage numeraire drop out:

$$\Delta\log C^u = (I-A^{uu})^{-1}\left(\Delta\log\mu^{uu} - \frac{\Delta\log N^u}{\sigma-1} + \Delta\log\mathcal{V}^u\right), \tag{32}$$

with $I - A^{uu}$ invertible since $\rho(A^{uu}) < 1$ when cost shares sum to less than one in each upstream sector.

Stacking and the linear-systems representation. Let

$$p_u := \operatorname{vec}\left([\,p^u_{vs'}\,]_{v,s'}\right) \in \mathbb{R}^{|\mathcal{S}|^2}, \quad p_r := \operatorname{vec}\left([\,p^r_{s's}\,]_{s',s}\right) \in \mathbb{R}^{|\mathcal{S}|^2}, \quad p_c := \begin{bmatrix} p^c_1 \\ \vdots \\ p^c_{|\mathcal{S}|} \end{bmatrix} \in \mathbb{R}^{|\mathcal{S}|},$$

where $p^u_{vs'} = \log P^u_{vs'}$, $p^r_{s's} = \log P^r_{s's}$, and $p^c_s = \log P_s$. Stack $p := \left[p_u;\, p_r;\, p_c\right] \in \mathbb{R}^{|\mathcal{S}|^2+|\mathcal{S}|^2+|\mathcal{S}|}$.

Define the buyer-specific wedge matrices

$$\ln\mathbf{M}_u = [\,\ln\mu^u_{vs'}\,]_{v,s'}, \qquad \ln\mathbf{M}_r = [\,\ln\mu^r_{s's}\,]_{s',s}, \qquad \ln\mathbf{M}_c = \begin{bmatrix} \ln\mu^r_1 \\ \vdots \\ \ln\mu^r_{|\mathcal{S}|} \end{bmatrix}.$$

Let

$$n_u := \frac{\log N^u}{\sigma-1}, \qquad n_r := \frac{\log N^r}{\varphi-1}, \qquad z_u := \log\widetilde{z}^u, \qquad z_r := \log\widetilde{z}^r, \qquad \Theta_u := \log\Theta^u, \qquad \Theta_r := \log\Theta^r.$$

The stacked system is

$$p = Ap + d, \tag{33}$$

with blocks

$$A = \begin{bmatrix} A_{uu} & 0 & 0 \\ A_{ru} & 0 & 0 \\ 0 & A_{cr} & 0 \end{bmatrix}, \qquad d := \begin{bmatrix} d_{uu} \\ d_{ru} \\ d_{cu} \end{bmatrix},$$

and nonzero blocks

$$\begin{aligned}
&\text{Upstream–upstream:}\quad A_{uu}^{(s',v)} = (1-\alpha_{s'}^{u})\,\theta_{s'v}^{u}\,(\mathbf{1}_{|\mathcal{S}|}e_{s'}^{\top}) \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{S}|},\\
&\text{Retail–upstream:}\quad \omega_{s'}^{u} := (1-\alpha_{s'}^{u})\,[\,\theta_{s'1}^{u}\ \cdots\ \theta_{s'|\mathcal{S}|}^{u}\,],\quad A_{ru} = \Big[\,\mathbf{1}_{|\mathcal{S}|}\otimes\omega_{1}^{u}\ \cdots\ \mathbf{1}_{|\mathcal{S}|}\otimes\omega_{|\mathcal{S}|}^{u}\Big],\\
&\text{Consumer–retail:}\quad D_{s'} := \operatorname{diag}\Big((1-\alpha_{1}^{r})\theta_{1s'}^{r},\ldots,(1-\alpha_{|\mathcal{S}|}^{r})\theta_{|\mathcal{S}|\,s'}^{r}\Big),\quad A_{cr} = [\,D_1\ \cdots\ D_{|\mathcal{S}|}\,].
\end{aligned}$$

Affine terms:

$$d_{uu} = \operatorname{vec}(\ln\mathbf{M}_u)+(\Theta_u-z_u-n_u)\otimes\mathbf{1}_{|\mathcal{S}|},\quad d_{ru} = \operatorname{vec}(\ln\mathbf{M}_r)+(\Theta_r-z_r-n_r)\otimes\mathbf{1}_{|\mathcal{S}|},\quad d_{cu} = \ln\mathbf{M}_c+(\Theta_r-z_r-n_r).$$

Solving (33) forward:

$$p_u = (I-A_{uu})^{-1}\,d_{uu},\qquad p_r = A_{ru}(I-A_{uu})^{-1}\,d_{uu}+d_{ru},\qquad p_c = A_{cr}A_{ru}(I-A_{uu})^{-1}\,d_{uu}+A_{cr}d_{ru}+d_{cu}. \tag{34}$$

Let $\Omega_c := \theta^{\top}$ so that $\log P_Y = \Omega_c p_c$. Define the linear maps

$$\Lambda_{uu} := \Omega_c A_{cr} A_{ru}(I-A_{uu})^{-1},\qquad \Lambda_{ru} := \Omega_c A_{cr},\qquad \Lambda_c := \Omega_c, \tag{35}$$

which are $1\times|\mathcal{S}|^2$, $1\times|\mathcal{S}|^2$, and $1\times|\mathcal{S}|$, respectively. Then

$$-\log W = \log P_Y = \Lambda_{uu}\,d_{uu} + \Lambda_{ru}\,d_{ru} + \Lambda_c\,d_{cu}. \tag{36}$$

Aggregation of buyer-specific wedges and equivalence to exposure weights. Define the sector-level aggregates

$$\Big[\log\mu^{r}\Big]_{s'} := \frac{1}{\bar{B}_{s'}}\sum_{s}\theta_s(1-\alpha_s^{r})\theta_{ss'}^{r}\,\log\mu_{s's'}^{r},\qquad \Big[\log\mu^{uu}\Big]_{s'} := \sum_{v}A_{s'v}^{uu}\,\log\mu_{vs'}^{u}, \tag{37}$$

with $\bar{B} := \theta^{\top}B^{ru}$, so $\bar{B}_{s'} = \sum_s\theta_s(1-\alpha_s^{r})\theta_{ss'}^{r}$ is the retail–upstream interface weight for upstream sector s'.

Lemma 3 (Linear maps equal exposure-weighted aggregates). *For any perturbations of buyer-specific wedges,*

$$\Lambda_{ru}\ \operatorname{vec}(\Delta\log\mathbf{M}_r) = \tilde{\lambda}_{ru}\cdot\Delta\log\mu^{r},\qquad \Lambda_{uu}\ \operatorname{vec}(\Delta\log\mathbf{M}_u) = \tilde{\lambda}_{u}\cdot\Delta\log\mu^{uu}.$$

Proof. *For $\Lambda_{ru} = \Omega_c A_{cr}$, the block-column of A_{cr} associated with upstream sector s' is $D_{s'} = \mathrm{diag}((1-\alpha_s^r)\theta_{ss'}^r)_s$. Left-multiplying by $\Omega_c = \theta^\top$ collapses $D_{s'}$ to $\bar{B}_{s'}$. Hence*

$$\Lambda_{ru}\ \mathrm{vec}(\Delta\log\mathbf{M}_r) = \sum_{s'} \bar{B}_{s'}\left(\frac{\sum_s \theta_s(1-\alpha_s^r)\theta_{ss'}^r\,\Delta\log\mu_{s's}^r}{\bar{B}_{s'}}\right) = \sum_{s'}\tilde{\lambda}_{ru,s'}\,\Delta\log\mu_{s'}^r.$$

For Λ_{uu}, A_{ru} selects rows of p_u with weights ω^u, $(I-A_{uu})^{-1}$ propagates upstream linkages, and $\Omega_c A_{cr}$ yields exactly $\tilde{\lambda}_u$ from (20). *Aggregation with A^{uu} produces $\Delta\log\mu^{uu}$ as in* (37). □

Proof of Proposition 2. Using (34)–(36), expand $\log P_Y$ into markup, variety, and selection components. Take differences across equilibria; technology constants and the wage drop out. Apply Lemma 3 to replace $\Lambda_{ru}\,\mathrm{vec}(\Delta\log\mathbf{M}_r)$ and $\Lambda_{uu}\,\mathrm{vec}(\Delta\log\mathbf{M}_u)$ with $\tilde{\lambda}_{ru}\cdot\Delta\log\mu^r$ and $\tilde{\lambda}_u\cdot\Delta\log\mu^{uu}$, respectively. Finally, use $\Delta\log W = -\Delta\log P_Y$ and the definitions $n_r = \log N^r/(\varphi-1)$ and $n_u = \log N^u/(\sigma-1)$ to obtain exactly 2, with selection terms loading via $+\Lambda_{uu}\Delta\log\mathcal{V}^u$ and $+\Lambda_c\Delta\log\mathcal{V}$ in $\Delta\log P_Y$, hence with minus signs in $\Delta\log W$. This yields the stated intensive, variety, and selection groupings. □

Selection invariance. The selection terms are

$$\mathcal{V}_s = \left(\frac{1}{N_s^r}\int z^{\varphi_s-1}\,d\nu_{rs}(z)\right)^{-\frac{1}{\varphi_s-1}}, \qquad \mathcal{V}_{s'}^u = \left(\frac{1}{N_{s'}^u}\int z^{\sigma_{s'}-1}\,d\nu_{us'}(z)\right)^{-\frac{1}{\sigma_{s'}-1}}.$$

If the productivity composition within active sets is invariant across the comparison (for example, Pareto tails with common lower bound and unchanged truncation rule), then $\Delta\log\mathcal{V}_s = \Delta\log\mathcal{V}_{s'}^u = 0$ and selection drops out. Otherwise, selection loads with the same exposures θ and $\tilde{\lambda}_u$ as the corresponding variety terms.

C.7 Equilibrium existence and uniqueness

Roadmap. The decentralized equilibrium with two–part tariffs is pinned in six linked steps. S1 fixes the composition of retail labor from final demand. S2 maps retail labor into upstream labor through a linear network that depends on cost shares and buyer–specific μ markups. S3 stacks upstream free entry in labor units, yielding a linear relation between sectoral labor and upstream entry flows. S4 composes S2 and S3 and adds retail free entry, so entry flows on both layers are linear in retail labor. S5 imposes aggregate labor clearing on the retail ray, reducing the problem to one scalar t with a unique solution. S6 solves the price–cost block (log indices and wage) from linear CES/Cobb–Douglas relations; the coefficient matrix is a contraction. Under S1-S6, we can show equilibrium existances and uniqueness, where, quantities, masses, and prices are all uniquely determined.

Sectors are indexed by $s, m \in \{1, \dots, S\}$. Layer $\ell \in \{u, r\}$ denotes upstream or retail. Labor shares $\alpha_s^\ell \in (0, 1)$; materials shares sum to one by buyer: $\sum_m \theta_{u,m,\cdot} = 1$ for upstream buyers and $\sum_m \theta_{r,\cdot,m} = 1$ for retail buyers. Elasticities $\sigma_m > 1$ (upstream) and $\varphi_s > 1$ (retail). Final demand across retail sectors is Cobb–Douglas with weights $\boldsymbol{\theta}_c = (\theta_{c,1}, \dots, \theta_{c,S})$, $\mathbf{1}^\top \boldsymbol{\theta}_c = 1$. Buyer–specific per–unit markups are denoted by μ: $\mu_{s,m}^{u \to u}$ (upstream buyer s from upstream seller m), $\mu_{s,m}^{u \to r}$ (retail buyer s from upstream seller m), and μ_s^r (retail to final consumer in sector s). Entry costs $c_{e,s}^\ell > 0$; exit rates $\delta_s^\ell \in [0, 1)$. Masses N_s^ℓ, entry flows $e_s^\ell = (1 - \delta_s^\ell) N_s^\ell$. Stack vectors by sector: $\mathbf{L}_u, \mathbf{L}_r, \mathbf{e}_u, \mathbf{e}_r \in \mathbb{R}_+^S$. Lower case denotes logs (e.g. $w = \ln W$, $p = \ln P$).

Equilibrium conditions

1. Retail free entry: $\mathbb{E}[\Pi_s^r] = W\, c_{e,s}^r (1 - \delta_s^r)$ for all s.

2. Upstream free entry: $\mathbb{E}[\Pi_m^u] = W\, c_{e,m}^u (1 - \delta_m^u)$ for all m.

3. Average firm output meets demand within sector:

$$\widetilde{z}_{r,s}\, y_{r,s}(\widetilde{z}_{r,s}) = Y_{c,\cdot,s}\, N_{r,s}^{\frac{\varphi_s}{1-\varphi_s}}, \qquad \widetilde{z}_{u,m}\, y_{u,m}(\widetilde{z}_{u,m}) = D_{u,m}\, N_{u,m}^{\frac{\sigma_m}{1-\sigma_m}}.$$

4. CES unit–price indices for any buyer (b, s) from upstream seller m:

$$p_{b,s,m} = \ln \mu_{s,m}^{u \to b} + mc_{u,m} + \frac{1}{1 - \sigma_m}\, n_{u,m}.$$

   Upstream marginal cost: $mc_{u,m} = (\ln \Theta_{u,m} - \ln \widetilde{z}_{u,m}) + \alpha_m^u w + (1 - \alpha_m^u) \sum_j \theta_{u,m,j}\, p_{u,m,j}$. Retail marginal cost: $mc_{r,s} = (\ln \Theta_{r,s} - \ln \widetilde{z}_{r,s}) + \alpha_s^r w + (1 - \alpha_s^r) \sum_m \theta_{r,s,m}\, p_{r,s,m}$. Final–good index: $p_{c,s} = \ln \mu_s^r + mc_{r,s} + \frac{1}{1-\varphi_s} n_{r,s}$, $\sum_s \theta_{c,s} p_{c,s} = 0$.

5. Labor market clearing: $1 = \sum_s L_s^r + \sum_m L_m^u + \sum_s e_s^r + \sum_m e_m^u$.

S1. Final demand pins the retail composition (ray)

Cobb–Douglas final demand implies sectoral retail revenue shares equal preference weights. With CRS, $WL_s^r = \alpha_s^r\, \text{Revenue}_s^r$, so relative retail labor is fixed:

$$\bar{\mathbf{L}}_r \;\propto\; \left(\frac{\theta_{c,s}}{\alpha_s^r}\right)_{s=1}^{S} \gg 0, \qquad \mathbf{L}_r = t\, \bar{\mathbf{L}}_r, \quad t > 0.$$

Only the scalar t remains to be determined by aggregate labor clearing.

S2. Sectoral labor mapping with buyer–specific markups

At the sector level, CRS cost shares and CES demand under buyer–specific per–unit markups imply a linear system that maps retail labor into upstream labor and propagates upstream feedback:

$$\mathbf{L}_u = A\,\mathbf{L}_u + B\,\mathbf{L}_r,$$

with nonnegative $S \times S$ matrices

$$A_{s,m} = \alpha^u_s \,\frac{1-\alpha^u_m}{\alpha^u_m}\,\frac{\theta_{u,m,s}}{\mu^{u\to u}_{m,s}}, \qquad B_{s,r} = \alpha^u_s \,\frac{1-\alpha^r_r}{\alpha^r_r}\,\frac{\theta_{r,r,s}}{\mu^{u\to r}_{r,s}}.$$

Interpretation: a buyer in upstream sector s allocates the share $(1-\alpha^u_m)\theta_{u,m,s}$ of revenue to materials from upstream seller m, scaled by the buyer–specific markup faced on that link; the labor anchor α^u_s converts revenue to labor. Likewise for retail exposure B. Assume the spectral condition

$$\rho(A) < 1 \quad \text{(sufficient: each row sum of } A \text{ is} < 1\text{)},$$

so the Neumann series converges and the total upstream requirements per unit of retail labor are

$$\Psi := (I - A)^{-1}B \;\geq 0, \qquad \mathbf{L}_u = \Psi\,\mathbf{L}_r.$$

Higher per–unit markups weaken links (prices are higher), reducing the corresponding entries of A and B and, through Ψ, the upstream labor implied by a given $\mathbf{L}_r$.

S3. Upstream free entry in labor units (stacked)

Two–part tariffs imply that an upstream entrant's expected period profit equals a constant fraction of buyer spending (variable margin) plus net flat–fee transfers (received from buyers minus paid to own suppliers). Dividing by W makes profits linear in buyer revenues expressed in labor units:

$$(G^{uu} - \Pi_u)\,\mathbf{L}_u \;+\; G^{ru}\,\mathbf{L}_r \;=\; C_u\,\mathbf{e}_u,$$

where $G^{uu}, G^{ru} \geq 0$ are flow–weight matrices determined by cost shares and the tariff schedule, $\Pi_u = \mathrm{diag}((G^{uu})^\top \mathbf{1})$ nets out intra–upstream transfers, $C_u = \mathrm{diag}(c^u_{e,s})$, and $\mathbf{e}_u$ stacks upstream entry flows. The left side converts sectoral labor into expected upstream profits by sector; the right side is entry cost (in labor units) times the entrant flow, as required by free entry.

S4. Entry maps as linear functions of retail labor

Compose S2 into S3:

$$\mathbf{e}_u = C_u^{-1}\Big[(G^{uu} - \Pi_u)\Psi + G^{ru}\Big]\mathbf{L}_r \;=\; \chi_u\,\mathbf{L}_r, \qquad \chi_u \geq 0.$$

Retail free entry implies $e_s^r = (1-\delta_s^r)N_s^r = (1-\delta_s^r)L_s^r/l_s^r$. Let $\Gamma_r = \operatorname{diag}((1-\delta_s^r)/l_s^r) > 0$. Then

$$\mathbf{e}_r = \Gamma_r\,\mathbf{L}_r.$$

Thus, once the retail composition $\bar{\mathbf{L}}_r$ is fixed by S1, both upstream and retail entry flows move linearly with the scale t along that ray.

S5. Labor clearing on the retail ray

Total labor equals production labor plus entry labor:

$$1 = \mathbf{1}^\top\Big(\mathbf{L}_r + \mathbf{L}_u + \mathbf{e}_r + \mathbf{e}_u\Big) = \mathbf{1}^\top\Big[\Big(I + \Psi + \Gamma_r + \chi_u\Big)\,\mathbf{L}_r\Big].$$

On the ray $\mathbf{L}_r = t\,\bar{\mathbf{L}}_r$, define

$$\Xi(t) := \mathbf{1}^\top\Big(I + \Psi + \Gamma_r + \chi_u\Big)(t\,\bar{\mathbf{L}}_r).$$

Because $(I + \Psi + \Gamma_r + \chi_u)\bar{\mathbf{L}}_r \gg 0$, Ξ is continuous, strictly increasing, $\Xi(0) = 0$, and $\Xi(t) \to \infty$ as $t \to \infty$. There exists a unique $t^\star > 0$ such that $\Xi(t^\star) = 1$. This pins

$$\mathbf{L}_r^\star = t^\star\bar{\mathbf{L}}_r, \qquad \mathbf{L}_u^\star = \Psi\,\mathbf{L}_r^\star, \qquad \mathbf{e}_u^\star = \chi_u\,\mathbf{L}_r^\star, \qquad \mathbf{e}_r^\star = \Gamma_r\,\mathbf{L}_r^\star,$$

and, hence, sectoral masses $N_s^\ell = e_s^\ell/(1-\delta_s^\ell)$.

S6. Price/cost block under wage normalization

The nominal wage is the numeraire. Write the nominal wage as ω and set $\omega = 1$, so its log is $w = \ln\omega = 0$. This step solves for sectoral price indices conditional on the masses from S1–S5, and then computes welfare.

For each buyer layer–sector pair $(b, s) \in \{u, r\} \times \{1, \ldots, S\}$ and upstream seller m, the CES unit price index satisfies

$$p_{b,s,m} \;=\; \ln\mu_{m,s}^{u\to b} \;+\; mc_{u,m} \;+\; \frac{1}{1-\sigma_m}\,n_{u,m}, \tag{P*}$$

where $p_{b,s,m} = \ln P_{b,s,m}$, $n_{u,m} = \ln N_m^u$, and $\mu_{m,s}^{u\to b}$ is the buyer–specific markup applied by upstream sector m when selling to buyer (b, s). Upstream marginal costs in logs are

$$mc_{u,m} \;=\; \left(\ln\Theta_{u,m} - \ln\widetilde{z}_{u,m}\right) \;+\; (1-\alpha_m^u)\sum_{j=1}^{S}\theta_{u,m,j}\,p_{u,m,j} \tag{MC–U$\star$}$$

and retail marginal costs are

$$mc_{r,s} \;=\; \left(\ln\Theta_{r,s} - \ln\widetilde{z}_{r,s}\right) \;+\; (1-\alpha_s^r)\sum_{m=1}^{S}\theta_{r,s,m}\,p_{r,s,m} \tag{MC–R$\star$}$$

Final–good price indices by retail sector are

$$p_{c,s} \;=\; \ln\mu_s^r \;+\; mc_{r,s} \;+\; \frac{1}{1-\varphi_s}\,n_{r,s}, \tag{PC$\star$}$$

where $n_{r,s} = \ln N_s^r$ and μ_s^r is the retail–to–consumer markup in sector s. For notational economy define the technology–selection constants

$$\xi_{u,m} := \ln\Theta_{u,m} - \ln\widetilde{z}_{u,m}, \qquad \xi_{r,s} := \ln\Theta_{r,s} - \ln\widetilde{z}_{r,s}.$$

Collect upstream–to–upstream indices by seller into the vector $p_u^u \in \mathbb{R}^S$ and define the nonnegative matrix

$$B_u \;:=\; \mathrm{diag}(1-\alpha^u)\,\Theta_u \;\in\; \mathbb{R}^{S\times S}, \qquad \|B_u\|_\infty = \max_m(1-\alpha_m^u) < 1.$$

Using (P$\star$) and (MC–U$\star$) in seller–stacked form,

$$(I-B_u)\,p_u^u \;=\; \ln\mu^{u\to u} \;+\; \xi_u \;+\; \frac{1}{1-\sigma}\,n_u, \qquad p_u^r \;=\; p_u^u \;+\; \ln\!\left(\mu^{u\to r}\oslash\mu^{u\to u}\right), \tag{U$\star$}$$

where $\oslash$ denotes componentwise division and n_u, ξ_u, $\ln\mu^{u\to\ell} \in \mathbb{R}^S$ are the seller–indexed vectors. Because $\|B_u\|_\infty < 1$, $I - B_u$ is invertible and the upstream fixed point is unique.

Retail sector indices follow from (P$\star$) and (MC–R$\star$) as

$$p_r[s] \;=\; \ln\mu_s^r \;+\; \xi_{r,s} \;+\; (1-\alpha_s^r)\sum_{m=1}^{S}\theta_{r,s,m}\,p_u^r[m] \;+\; \frac{1}{1-\varphi_s}\,n_{r,s}. \tag{R$\star$}$$

With the wage fixed at one, real income equals the inverse final price index. Writing the final–demand weights as $\boldsymbol{\theta}_c = (\theta_{c,1},\ldots,\theta_{c,S})^\top$ and $\Omega_c := \boldsymbol{\theta}_c^\top$, welfare is

$$\mathcal{W} \;=\; -\,\Omega_c\,p_c \;=\; -\sum_{s=1}^{S}\theta_{c,s}\,p_{c,s}, \tag{W$\star$}$$

with p_c obtained from (PC*) using p_u^u, p_u^r, and p_r computed above.

Existence and uniqueness of the price block under wage normalization. The upstream map $p_u^u \mapsto (I - B_u)^{-1}\Big(\ln \mu^{u \to u} + \xi_u + \frac{1}{1-\sigma} n_u\Big)$ is well defined and single–valued because $\|B_u\|_\infty < 1$ implies $\rho(B_u) < 1$ and hence $I - B_u$ is invertible. Given p_u^u, the transformation to p_u^r is an additive constant shift governed by buyer–specific markup gaps, therefore also unique. Retail indices are then affine in p_u^r, and final–good indices are affine in mc_r and n_r, so both are uniquely pinned. The entire price/cost block is linear in the unknown indices once masses n_u, n_r are taken from S1–S5, and the numeraire has already fixed the absolute price level by setting $\omega = 1$. No additional normalization is required, and uniqueness follows from the upstream contraction and forward substitution in the remaining equations.

Existence and uniqueness with alternative price regimes Nonlinear pricing with two–part tariffs retains buyer–specific per–unit markups $\mu_{m,s}^{u \to b}$ and permits flat fees that do not enter unit price indices. Steps S1–S5 are unchanged, since they rely on constant–returns shares and free entry in labor units. In S6, the only difference is in the constants of the upstream and retail equations through $\ln \mu$; the upstream coefficient matrix $I - B_u$ and the retail aggregation weights remain the same. Because $\|B_u\|_\infty < 1$ still holds, the upstream fixed point is unique and the rest of the system follows uniquely by forward substitution. Existence and uniqueness of the full equilibrium therefore carry over under nonlinear pricing with wage normalization.

Planner–implemented marginal–cost pricing sets effective per–unit markups to one, $\mu^{u \to u} \equiv \mu^{u \to r} \equiv 1$. The price block strictly simplifies: equation (U*) becomes $(I - B_u)\, p_u^u = \xi_u + \frac{1}{1-\sigma} n_u$ with the same B_u, and $p_u^r = p_u^u$. Retail indices follow from (R*) with $\ln \mu^r$ set to zero if the planner eliminates retail markups. The contraction property is preserved, so prices are uniquely pinned under $\omega = 1$. Regarding the quantity side, either the planner chooses masses directly by appropriate entry transfers, in which case S1–S5 run with masses treated as given, or the planner supports decentralized free entry with linear subsidies that preserve the linear mapping from labor to entry in labor units. In both implementations the one–dimensional labor–clearing step and the linear price block continue to deliver a unique equilibrium under the same spectral condition on upstream feedback, now in a system with strictly simpler coefficients.

D Parameter Calibration and Estimation

D.1 Labor Output Elasticity

In the Chilean firm data, we observe expenditures on three cost components: labor (wL), capital services (rK), and intermediate materials expenditure (M). Since we assume that firms minimize costs, the cost share mapping is valid and delivers a theoretically consistent estimate of the model's labor elasticity. In contrast, the model abstracts from capital and specifies production as a function of labor and materials only. This creates a mismatch.

To resolve this, we construct a bridge between model and data by bundling labor and capital into a single composite of "non–material" inputs primary inputs. In the data, this bundle is measured as $wL + rK$. We then define the labor share parameter as labor's plus capital proportion in total costs, or conversely, the non–material bundle:

$$\alpha_i \; = \; 1 - \frac{\sum_j p_{ji} m_{ji}}{w_i L_i + rK_i + \sum_j p_{ji} m_{ji}}.$$

Because upstream firms may charge two–part tariffs, total payments equal a flat fee plus variable payments: $TC_i = F_i + VC_i$, with $VC_i := w_i L_i + r_i K_i + \sum_j p_{ji} m_{ji}$. The mapping above holds for variable cost. Total-cost shares equal variable-cost shares scaled by $(1 - esc)$, where $esc_i := F_i/(F_i + VC_i)$. For large buyers (high VC_i), esc_i is small, so total-cost shares are well approximated by variable-cost shares, and we use α_i as the labor (primary) output elasticity.

We keep firms above the 75th percentile of revenue in each year, winsorize the share α_i at the 1st and 99th percentiles, and compute α_{s_ℓ} for each firm type ℓ and each 6-digit sector. We do this separately by year over 2005–2022 and then average across years. Table A5 reports sectoral means by 1-digit sector and firm type.

Table A5: Labor Shares by Sector (mean)

Sector	Retailers	Upstream	Sector mean
Agriculture	0.43	0.41	0.42
Mining	0.25	0.32	0.29
Manufacturing	0.39	0.42	0.41
Utilities	0.37	0.38	0.38
Construction	0.48	0.42	0.45
Retail and Wholesale	0.37	0.31	0.34
Transport and ICTs	0.55	0.47	0.51
Financial Services	0.58	0.62	0.60
Real Estate Services	0.66	0.53	0.59
Business Services	0.72	0.65	0.69
Personal Services	0.71	0.57	0.64
Type mean	0.50	0.46	0.48

D.2 Input-output and Output Elasticities

Input-output Elasticities. Using firm–to–firm transactions, we recover buyer–facing variable expenditure shares on upstream sectors. For buyer firm i of type $\ell \in \{r, u\}$ in 6–digit sector j and upstream seller sector q, define

$$\theta^{\ell}_{ijq} := \frac{\sum_{k \in \mathcal{U}_q} p_{ik}\, m_{ik}}{\sum_{r \in \mathcal{S}_U} \sum_{k \in \mathcal{U}_r} p_{ik}\, m_{ik}}, \qquad \sum_q \theta^{\ell}_{ijq} = 1,$$

where p_{ik} is the buyer–facing marginal price for upstream variety k and m_{ik} the corresponding quantity. As with labor, two–part tariffs imply total payments $TC_i = F_i + VC_i$, so θ^{ℓ}_{ijq} maps to variable materials shares; for large buyers (high VC_i), the flat–fee share F_i/TC_i is small and total–cost shares are well approximated by variable shares.

We compute firm–level θ^{ℓ}_{ijq} separately for retailers ($\ell = r$) and upstream buyers ($\ell = u$), retain firms above the 75th percentile of revenue in each year, aggregate from 6–digit industries to the buyer's 1–digit sector j by simple averaging within year, and then average across 2005–2022. Rows sum to one up to rounding. Table A6 reports retailer buyers; Table A7 reports upstream buyers.

Table A6: Input-output Elasticities by Retailers as Buyers

Buyer \ Seller	Agr.	Min.	Man.	Uti.	Cons.	R. & W.	T. & ICTs	F. Serv.	RE. Serv.	B. Serv.	P. Serv.
Agriculture	0.25	0.00	0.21	0.02	0.03	0.32	0.05	0.07	0.00	0.04	0.00
Mining	0.00	0.04	0.19	0.06	0.15	0.30	0.07	0.02	0.00	0.17	0.00
Manufacturing	0.13	0.02	0.35	0.02	0.03	0.25	0.11	0.03	0.00	0.06	0.00
Utilities	0.07	0.01	0.18	0.03	0.03	0.26	0.17	0.05	0.00	0.20	0.00
Construction	0.10	0.00	0.10	0.02	0.22	0.24	0.15	0.03	0.00	0.14	0.00
Retail and Wholesale	0.16	0.01	0.24	0.01	0.02	0.34	0.08	0.05	0.00	0.09	0.00
Transport and ICTs	0.07	0.01	0.14	0.02	0.03	0.24	0.19	0.04	0.00	0.26	0.00
Financial Services	0.08	0.00	0.12	0.01	0.01	0.22	0.06	0.15	0.01	0.33	0.00
Real Estate Services	0.03	0.00	0.12	0.01	0.02	0.30	0.04	0.06	0.05	0.37	0.00
Business Services	0.07	0.00	0.13	0.01	0.01	0.22	0.09	0.06	0.00	0.41	0.00
Personal Services	0.07	0.00	0.17	0.02	0.02	0.25	0.07	0.08	0.00	0.33	0.01

Table A7: Input-output Elasticities by Upstream Firms as Buyers

Buyer \ Seller	Agr.	Min.	Man.	Uti.	Cons.	R. & W.	T. & ICTs	F. Serv.	RE. Serv.	B. Serv.	P. Serv.
Agriculture	0.26	0.00	0.12	0.02	0.04	0.29	0.10	0.06	0.00	0.10	0.00
Mining	0.01	0.07	0.39	0.05	0.06	0.13	0.11	0.03	0.00	0.15	0.00
Manufacturing	0.08	0.02	0.49	0.03	0.02	0.15	0.09	0.02	0.00	0.10	0.00
Utilities	0.06	0.02	0.18	0.07	0.03	0.18	0.15	0.04	0.00	0.27	0.00
Construction	0.07	0.00	0.14	0.03	0.30	0.18	0.12	0.03	0.00	0.13	0.00
Retail and Wholesale	0.12	0.01	0.27	0.01	0.02	0.38	0.07	0.03	0.00	0.10	0.00
Transport and ICTs	0.06	0.02	0.14	0.02	0.04	0.21	0.22	0.03	0.00	0.26	0.00
Financial Services	0.05	0.00	0.12	0.02	0.01	0.20	0.07	0.12	0.01	0.41	0.00
Real Estate Services	0.03	0.00	0.11	0.01	0.02	0.27	0.04	0.04	0.06	0.41	0.00
Business Services	0.07	0.00	0.13	0.01	0.01	0.23	0.09	0.05	0.00	0.40	0.00
Personal Services	0.06	0.00	0.15	0.03	0.02	0.21	0.07	0.11	0.00	0.33	0.01

Cobb–Douglas Output Elasticities. Under a Cobb–Douglas aggregation of final demand across retail sectors with within–sector CES over firm varieties, sectoral budget shares are constant and equal to the Cobb–Douglas output elasticities θ_s. Because retail–to–final transactions are linear in our setting, observed retail revenues identify each sector's final expenditure. We therefore estimate θ_s as the sector's share of aggregate retail expenditure. Operationally, within each year we restrict to retailers above the 75th percentile of revenue, compute firm–level revenue shares, take the simple mean within sector s, and then average across years (2005–2022). This unweighted mean provides a transparent proxy for the revenue–weighted sector share; we also verify robustness using revenue weights.

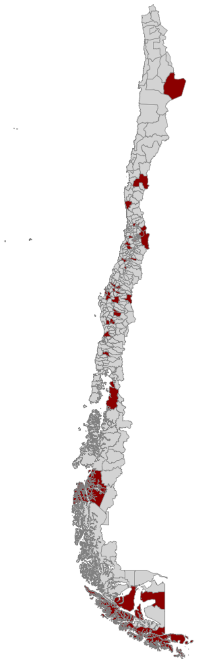
Table A8: Cobb–Douglas Output Elasticities by Retail Sector

Sector	θ_s
Agriculture	0.0446
Mining	0.0085
Manufacturing	0.1318
Utilities	0.0505
Construction	0.1521
Retail and Wholesale	0.2768
Transport and ICTs	0.0979
Financial Services	0.1132
Real Estate Services	0.0152
Business Services	0.0911
Personal Services	0.0183

D.3 Upstream Materials Elasticity of Substitution

We estimate the elasticity of substitution ($\sigma_{u'}$) across upstream input varieties, exploiting a one-off, municipality-level cost shock from Chile's March 2020 COVID-19 lockdowns. March 2020 marks the first registered COVID-19 cases in Chile, making the shock unexpected. Figure A3 shows the spatial heterogeneity of the early lockdowns (municipalities in red were under lockdown in March 2020).

Figure A3: Distribution of early COVID-19 lockdowns in Chile



Buyers aggregate upstream inputs from seller sector u' with a CES technology. For buyer (i, s), let u^* denote the *largest* pre-shock supplier (by 2019 expenditure). For any $u \in \mathcal{U}_{u'}$,

$$\log\left(\frac{m_{isut}}{m_{isu^*t}}\right) = -\sigma_{u'} \log\left(\frac{p_{isut}}{p_{isu^*t}}\right).$$

Data come from the Chilean Internal Revenue Service (SII): monthly firm-to-firm transactions, 2019–2021, including product descriptions, quantities, prices, firm identifiers, and locations. We match buyers to their upstream suppliers, observe the universe of intermediate-input purchases, and track geographic lockdown exposure. We focus on intermediate-input links $(i, s) \times u \in \mathcal{U}_{u'}$ and identify u^* for each (i, s) as the 2019 top supplier by value.

Define a binary instrument $Z_{isu} = 1$ if the main supplier u^* was located in a municipality under lockdown in March 2020, and 0 otherwise (noting Z_{isu} is determined by u^* and is therefore constant across u for a given (i, s)). This captures a plausibly exogenous increase in u^*'s marginal cost (and price) relative to other inputs, inducing substitution.

To ensure unit prices reflect marginal (allocative) prices, the estimation sample is restricted to large buyers (above the 90th percentile of average annual sales in 2019–2021). The core exclusion is that the shock affects the buyer only through u^*'s relative cost. To mitigate alternative channels, we impose:

1. **Buyer location**: the buyer is not in a locked municipality.

2. **Client base**: the buyer's customers are not in locked municipalities.

3. **Input scope**: no other input used by the buyer was sourced from a locked municipality.

We use 12-month log differences in relative prices and quantities to remove time-invariant buyer–supplier heterogeneity and seasonality, and we include buyer 6-digit sector fixed effects γ_s to absorb sector-level shocks common to all inputs. Standard errors are clustered at the buyer level.

For each seller sector u', we estimate a separate 2SLS on $(i, s) \times u \in \mathcal{U}_{u'}$:

$$\textbf{First stage:} \quad \Delta_{12} \log\left(\frac{p_{isut}}{p_{isu^*t}}\right) = \beta_0 + \beta_1 Z_{isu} + \gamma_s + \nu_{isut},$$

$$\textbf{Second stage:} \quad \Delta_{12} \log\left(\frac{m_{isut}}{m_{isu^*t}}\right) = \beta_{u'}\, \Delta_{12} \widehat{\log\left(\frac{p_{isut}}{p_{isu^*t}}\right)} + \gamma_s + \varepsilon_{isut},$$

where $\beta_{u'} = -\sigma_{u'}$. The 12-month horizon targets the long-run notion of substitution relevant for counterfactuals.

The resulting $\hat{\sigma}_{u'}$, one per upstream seller sector, feed directly into the structural model, governing the curvature of sectoral revenue functions and the strength of intensive-margin reallocation. Estimates are reported in Table A9.

Table A9: Estimated Elasticicities of Substitution by Seller Sector

Sector	$\sigma_{u'}$	SE	1^{st} Stage F stat.	Obs.
Agriculture	2.59	(1.35)	10.24	4,387
Manufacturing	3.41	(0.84)	16.37	186,912
Construction	1.45	(0.42)	7.36	6,062
Retail and Wholesale	3.80	(0.39)	94.08	680,985
Transport and ICTs	5.07	(2.22)	25.19	24,054
Financial Services	3.09	(1.56)	9.35	3,631
Business Services	5.21	(2.02)	17.55	4,514
Personal Services	6.69	(3.37)	13.29	7,579
All sectors	3.04	(1.12)	149.87	918,124

Three sectors, Mining, Utilities, and Real Estate Services, yield $\hat{\sigma}_{u'} < 1$; for these, we set $\sigma_{u'}$ to the minimum estimate above one, 1.45, for use in the model quantification.

D.4 Exit Rates and Entry Costs

We estimate, from administrative firm microdata for Chile, two parameters: (i) the annual exit hazard δ_{s_ℓ} for each sector–firm type (s,ℓ), and (ii) the entry cost c_{e,s_ℓ} measured in units of yearly firm-level wages. The calibration exploits the availability of firm-level accounting profits and wages. We consider a one-year period and a panel of firms indexed by i, years by t, sectors by s, and firm types by $\ell \in \{\text{upstream}, \text{retail}\}$. We compute both parameters at 6-digit sector granularity (626 sectors). Let $\mathcal{T} = \{2005, \ldots, 2022\}$ denote the estimation window.

Exit Rate. For each firm-year observation (i,t), define an indicator for one-year survival

$$\text{surv}_{i,t} := \nVdash\{\exists \text{ an observation of firm } i \text{ in year } t+1\}.$$

At the sector–firm type–year level, let $\text{active}_{s_\ell,t}$ be the number of firms observed in (s_ℓ,t) and $\text{survivors}_{s_\ell,t} = \sum_{i\in(s_\ell,t)} \text{surv}_{i,t}$. Denote $N_{s_\ell,t} = \text{active}_{s_\ell,t}$. The exit rate in year t is

$$\delta_{s_\ell,t} \;=\; 1 - \frac{\text{survivors}_{s_\ell,t}}{N_{s_\ell,t}}.$$

We then fix the cell-level hazard as the window average

$$\delta_{s_\ell} \;=\; \frac{1}{|\mathcal{T}|}\sum_{t\in\mathcal{T}} \delta_{s_\ell,t}.$$

Table A10 reports 1-digit sector means of the 6-digit estimates.

Table A10: Exit Rates (δ) by Sector (Means)

Sector	Retailers	Upstream	Sector mean
Agriculture	0.090	0.086	0.088
Mining	0.084	0.093	0.088
Manufacturing	0.093	0.071	0.082
Utilities	0.070	0.064	0.067
Construction	0.140	0.110	0.125
Retail and Wholesale	0.103	0.076	0.089
Transport and ICTs	0.088	0.093	0.091
Financial Services	0.101	0.062	0.081
Real Estate Services	0.115	0.099	0.107
Business Services	0.099	0.077	0.088
Personal Services	0.093	0.090	0.092
Type mean	0.098	0.084	0.091

Entry Cost. We have access to data on firm-level yearly revenue, labor headcounts, wage-bill expenditure, material expenditure, and capital stock. We build the real user cost of capital using publicly available data.[25] Using these data, we construct yearly firm-level profits $\Pi_{i,t}$ for 2005–2022 and compute the average wage per worker-year

$$w_{i,t} \equiv \frac{\text{wagebill}_{i,t}}{\text{employees}_{i,t}}.$$

We define the per-active-firm annual profit in (s_ℓ, t) as

$$\bar{\Pi}_{s_\ell,t} = \frac{1}{N_{s_\ell,t}} \sum_{i \in (s_\ell,t)} \Pi_{i,t}, \qquad N_{s_\ell,t} = \text{active}_{s_\ell,t}.$$

Let the sector–firm type wage per worker-year be the headcount-weighted mean

$$w_{s_\ell,t} = \frac{\sum_{i \in (s_\ell,t)} w_{i,t} \cdot \text{employees}_{i,t}}{\sum_{i \in (s_\ell,t)} \text{employees}_{i,t}}.$$

[25] We use the 10-year government bond interest rate minus expected inflation plus the external financing premium. We use the capital depreciation rate from the LA-KLEMS database. For reference, the average government bond interest rate over 2005–2022 is 5.74%, expected inflation is 4.6%, the external financing premium is 110 basis points, and the average capital depreciation rate is 5%.

We set w_{s_ℓ} and $\bar{\Pi}_{s_\ell}$ as averages over $\mathcal{T}$.

In our model, flat fees can redistribute profits across firm types. Calibration must therefore use the profit that accrues to the owner of the firm that pays entry. Under steady state with i.i.d. per-period profits and exogenous exit hazard δ_{s_ℓ}, the expected present value (PV) of a surviving firm is

$$\mathrm{PV}_{s_\ell} \;=\; \frac{\bar{\Pi}_{s_\ell}}{1-\beta(1-\delta_{s_\ell})},$$

where $\beta = 1/(1+r)$ is the annual real discount factor and r is the annual real rate. As only a fraction $p^{\text{succ}}_{s_\ell} \in (0,1]$ of firms have positive profits, we set $p^{\text{succ}}_{s_\ell}$ to the share of positive-profit firms in (s_ℓ). The free-entry condition is

$$w_{s_\ell}\, c_{e,s_\ell} \;=\; p^{\text{succ}}_{s_\ell} \cdot \mathrm{PV}_{s_\ell} \quad\Longrightarrow\quad c_{e,s_\ell} \;=\; \frac{p^{\text{succ}}_{s_\ell}}{w_{s_\ell}} \cdot \frac{\bar{\Pi}_{s_\ell}}{1-\beta(1-\delta_{s_\ell})}\,.$$

We compute entry costs by firm type and 6-digit sector. Table A11 presents 1-digit sector averages. We also report the equivalent number of yearly wage bills (total yearly labor costs) implied by the entry cost.

Table A11: Entry Costs and Equivalent Yearly Wage-Bills by Sector

	Retailers		Upstream	
Sector	Entry cost c_e	Wage-bill eq.	Entry cost c_e	Wage-bill eq.
Agriculture	81.03	3.68	84.12	4.78
Mining	29212.81	43.99	177.12	7.20
Manufacturing	101.87	4.25	120.80	4.53
Utilities	700.66	14.15	306.11	5.50
Construction	109.72	7.78	109.05	4.18
Retail and Wholesale	63.92	6.06	83.61	5.13
Transport and ICTs	299.85	10.28	98.03	6.40
Financial Services	263.84	8.64	248.44	9.05
Real Estate Services	82.11	11.68	100.69	8.70
Business Services	82.91	5.76	125.21	3.11
Personal Services	127.87	4.56	94.76	4.57
Type mean	2829.69	10.98	140.72	5.74

Notes: Entry costs c_e are in the currency units used for calibration; "Wage-bill eq." reports multiples of the annual wage bill.

The mapping from profits to entry costs we use is dimensionally consistent (output units into labor units via w_{s_ℓ}) and directly comparable across sector–firm types. Compared to alternatives (e.g., inferring c_e from net-entry rates and size distributions or from structural Markov dynamics), this method is empirically simple, requires fewer auxiliary moments, and allows clean sectoral heterogeneity through the observed $\bar{\Pi}_{s_\ell}$ and w_{s_ℓ}.

D.5 Pareto Productivity Tails

Let $\ell \in \{u, r\}$ index the firm type (upstream, retail) and $s \in \{1, \ldots, 12\}$ index 1-digit sectors within each firm type. For firm i in (ℓ, s), let $L_i^{\ell s}$ be its number of workers. We assume a Pareto tail for $L^{\ell s}$:

$$\Pr\left(L^{\ell s} > l\right) \;=\; \left(\frac{L_{\min}^{\ell s}}{l}\right)^{\nu_s^\ell}, \qquad l \geq L_{\min}^{\ell s}, \quad \nu_s^\ell > 0,$$

equivalently, the density is $f_{L^{\ell s}}(l) = \nu_s^\ell \left(L_{\min}^{\ell s}\right)^{\nu_s^\ell} l^{-(\nu_s^\ell+1)}$. Given a threshold $L_{\min}^{\ell s}$ (baseline: firms with at least two employees), the closed-form MLE of the survival exponent is

$$\widehat{\nu_s^\ell} \;=\; \frac{n_{\ell s}}{\displaystyle\sum_{i:\, L_i^{\ell s} \geq L_{\min}^{\ell s}} \ln\left(\frac{L_i^{\ell s}}{L_{\min}^{\ell s}}\right)}, \qquad \mathrm{SE}\left(\widehat{\nu_s^\ell}\right) \approx \frac{\widehat{\nu_s^\ell}}{\sqrt{n_{\ell s}}},$$

where $n_{\ell s}$ is the number of tail observations.[26] We estimate ν_s^ℓ at the 1-digit sector for each layer and then map to productivity tails via the model's $l(z)$.

The model implies labor demand at productivity z:

$$l(z) \;=\; l(\tilde{z}) \left(\frac{z}{\tilde{z}}\right)^{\sigma-1}, \qquad \sigma > 1,$$

which is strictly increasing in z. Fix a firm type–sector pair (ℓ, s). Suppose productivity $Z^{\ell s}$ has a Pareto upper tail with survival exponent $\kappa_s^\ell > 0$:

$$\Pr\left(Z^{\ell s} > z\right) \;=\; \left(\frac{z_{\min}^{\ell s}}{z}\right)^{\kappa_s^\ell} \quad \text{for } z \geq z_{\min}^{\ell s}.$$

Let $L^{\ell s} = l(z^{\ell s})$ with $l(z) = l(\tilde{z})\,(z/\tilde{z})^{\sigma-1}$ and define $L_{\min}^{\ell s} = l(z_{\min}^{\ell s})$. Then $L^{\ell s}$ has a Pareto upper tail with survival exponent

$$\nu_s^\ell \;=\; \frac{\kappa_s^\ell}{\sigma - 1}, \qquad \Longleftrightarrow \qquad \kappa_s^\ell \;=\; (\sigma - 1)\,\nu_s^\ell.$$

[26] We use the continuous Pareto tail for integer L; in the upper tail this approximation is standard and accurate.

Therefore, given a labor Pareto tail ν_s^ℓ and the seller-sector elasticity of substitution $\sigma_{u'}$ for sector s, the productivity Pareto tail is pinned down by $\kappa_s^u = (\sigma_{u'} - 1)\nu_s^u$ for upstream firms and $\kappa_s^r = (\varphi_{s'} - 1)\nu_s^r$ for retailers firms . We estimate labor Pareto tails using the MLE above and show them in Table A12, together with the implied productivity tails using the estimated elasticities of substitution by sector.

Table A12: Labor and Implied Productivity Pareto Tails by Sector

	Retailers		Upstream		
Sector	ν_r	$\kappa_r = (\varphi_s - 1)\nu_r$	ν_u	$\kappa_u = (\sigma_{u'} - 1)\nu_u$	$\sigma_{u'}$
Agriculture	2.49	8.82	2.63	4.18	2.59
Mining	1.43	2.40	2.20	0.99	1.45
Manufacturing	2.66	8.58	2.15	5.18	3.41
Utilities	2.17	6.38	1.94	0.87	1.45
Construction	3.23	5.13	2.19	0.99	1.45
Retail and Wholesale	3.45	24.74	2.40	6.72	3.80
Transport and ICTs	2.20	2.32	3.04	12.37	5.07
Financial Services	2.55	1.02	2.26	4.72	3.09
Real Estate Services	4.36	3.59	3.03	1.36	1.45
Business Services	2.45	4.25	1.93	8.13	5.21
Personal Services	2.03	3.17	2.58	14.69	6.69

Notes: $\kappa = (\sigma_{u'} - 1)\nu$ uses seller-sector elasticities $\sigma_{u'}$ from Table A9. For Mining, Utilities, and Real Estate Services, we set $\sigma_{u'} = 1.45$ (minimum estimate above one).

D.6 Final-Consumer Elasticities of Substitution

We recover the elasticity of substitution faced by the representative final consumer across retail varieties within each retail sector s_r, denoted $\varphi_{s_r} > 1$. Under the Dixit–Stiglitz CES aggregator, the representative consumer allocates expenditure across differentiated retail varieties $j \in \mathcal{J}_{s_r}$ via

$$Q_{s_r} = \left(\sum_{j \in \mathcal{J}_{s_r}} q_j^{\frac{\varphi_{s_r} - 1}{\varphi_{s_r}}} \right)^{\frac{\varphi_{s_r}}{\varphi_{s_r} - 1}},$$

so the demand for variety j is isoelastic with elasticity φ_{s_r}. With linear pricing and monopolistic competition, the first-order condition yields the standard markup,

$$\mu_{s_r} \equiv \frac{p_j}{c_j} = \frac{\varphi_{s_r}}{\varphi_{s_r} - 1},$$

implying that the variable-profit share of revenue equals $1/\varphi_{s_r}$. Hence, for retailer j,

$$\Pi_j^{\text{var}} = \frac{1}{\varphi_{s_r}} R_j,$$

where R_j is sales revenue. Retailers also face a fixed operating cost $w_{s_r,t}F_{j,t}$, expressed in labor units $F_{j,t}$ and valued at the sectoral wage $w_{s_r,t}$. Accounting profits are

$$\Pi_{j,t} = \frac{1}{\varphi_{s_r}} R_{j,t} - w_{s_r,t}F_{j,t}.$$

Aggregating within sector s_r and year t gives

$$\sum_{j \in \mathcal{J}_{s_r}} \Pi_{j,t} = \frac{1}{\varphi_{s_r,t}} \sum_j R_{j,t} - w_{s_r,t} \sum_j F_{j,t},$$

which rearranges to the sector-year estimator

$$\varphi_{s_r,t} = \frac{\sum_j R_{j,t}}{w_{s_r,t} \sum_j F_{j,t} + \sum_j \Pi_{j,t}}.$$

Measurement. We restrict to firms above the 75th percentile of annual sales within each sector–year to improve measurement reliability and, given their larger input volumes, to minimize the influence of upstream flat fees on average costs. Revenues $R_{j,t}$ are observed directly as sales to final consumers and aggregated to the annual firm–sector level.

Table A13 presents the resulting estimates of yearly averages for the 205-2022 period of φ_{s_r} by 1-digit retail sector.

Table A13: Retailer Parameter φ_{s_r} by Sector

Sector	φ_{s_r}
Agriculture	4.54
Mining	2.68
Manufacturing	4.22
Utilities	3.94
Construction	2.59
Retail and Wholesale	8.17
Transport and ICTs	2.05
Financial Services	1.40
Real Estate Services	1.82
Business Services	2.73
Personal Services	2.56
Type mean	3.34

Notes: φ_{s_r} is computed from pooled sectoral sums of revenue, fixed costs (in labor units), and profits. The formula follows directly from the CES markup identity under linear pricing.

This approach recovers consumer-side elasticities from sectoral accounting identities under the structural model and requires no additional demand shifters or instruments. Because profits are systematically positive in our data, including $\sum_j \Pi_{j,t}$ in the denominator is essential for consistency with observed sectoral accounts.

E Quantification Material

E.1 Exposures to Final Consumption

This appendix reports the two exposure measures used to interpret the sectoral welfare opening. "SC share" is the sector's share in total supply–chain transaction value (not its share in final demand). The retail to upstream exposure indicates how strongly a marginal dollar of consumer spending reaches each upstream sector through retail purchases. The full upstream exposure lets that dollar continue circulating as upstream sectors buy from one another, capturing the total knock–on demand that ultimately lands in each upstream sector. For each measure we also report a normalized "Share" that sums to one across sectors, making the entries directly comparable.

Table A14: Final–consumption exposures by sector (levels and shares; three decimals)

Sector [SC share]	retail→upstream	Share(retail→upstream)	full upstream	Share(full upstream
Retail and Wholesale [32%]	0.170	0.319	0.462	0.36
Manufacturing [15%]	0.099	0.186	0.222	0.17
Transport and ICTs [10%]	0.051	0.096	0.172	0.13
Construction [8%]	0.065	0.121	0.107	0.08
Financial Services [18%]	0.046	0.087	0.090	0.07
Business Services [7%]	0.036	0.067	0.083	0.06
Agriculture [2%]	0.028	0.053	0.064	0.05
Utilities [3%]	0.019	0.036	0.031	0.02
Real Estate Services [1%]	0.010	0.018	0.017	0.01
Personal Services [2%]	0.006	0.012	0.011	0.00

Notes: "SC share" is the sector's share of supply–chain transaction value. The retail→upstream column tracks the immediate flow of consumer spending to upstream sectors via retail; the full upstream column adds all upstream–to–upstream rounds. Each "Share" column normalizes its exposure vector to sum to one. The mining sector (1%) has very small exposure in our data and is omitted for brevity.

The exposure map aligns closely with the sectoral welfare opening between nonlinear and linear pricing. Retail & Wholesale, Manufacturing, Transport/ICTs, and Construction carry the largest flows of final demand into the upstream network, so compressing markups where these exposures are high has the biggest aggregate payoff. Sectors with low exposures, such as Real Estate and Personal Services, contribute little to the aggregate gap even when their own wedges or masses move. In short, exposures indicate where reducing distortions yields the greatest system–wide gains: target the hubs through which most of the consumer dollar travels, and the welfare improvement is largest.